Comparing Action-Oriented Language in the Assessment of EFL Writing: An Action Research for Combining the First-Year Instruction of the National Curriculum and the International Baccalaureate (IB) Diploma Program in a Finnish High School

Kenneth Lai

Master's Thesis

English Studies
Faculty of Arts
University of Helsinki

January 2020

Tiedekunta – Fakultet – Faculty Arts	Koulutusohjelma – Utbildningsprogram – Degree Programme English Studies	
Opintosuunta – Studieinriktning – Study Track Applied Linguistics		
Tekijä – Författare – Author Kenneth Wenchen Lai		
Työn nimi – Arbetets titel – Title Comparing Action-Oriented Language in the Assessment of EFL Writing: An Action Research for Combining the First-Year Instruction of the National Curriculum and the International Baccalaureate (IB) Diploma Program in a Finnish High School		
Työn laji – Arbetets art – Level Master's	Aika – Datum – Month and year January 2021	Sivumäärä– Sidoantal – Number of pages 60pp + appendices (18pp)

Tiivistelmä – Referat – Abstract

This thesis compares the extent to which instruction and assessment in the Finnish National Curriculum (FNC) and International Baccalaureate (IB) at one Finnish high school align with the pedagogical approach to language instruction recommended in the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001, 2018). Comparison of the two curricula will be used to inform curriculum development at the school, where the aim is to combine pre–Diploma Program (DP)—i.e., instruction of first-year students who have been provisionally accepted to the DP program—and FNC instruction in the first year of high school.

Action-oriented language is envisioned in CEFR as a pedagogical approach that (1) treats language as a tool rather than an object for mastery and (2) recommends the instruction, assessment, and learning of the broad range of social contexts in which communication occurs. The first point draws largely on Focus on Form (FonF) approaches, developed in SLA research, while the second point draws largely on task-based language teaching (TBLT), developed in pedagogical research. While CEFR is regularly used today for benchmarking student language mastery, its uneven application in curriculum, course instruction, and course and exam assessment in the IB and FNC leave much to be desired, calling into question whether CEFR benchmarking can really be used for EFL students graduating from the IB and FNC.

This thesis uses a school in Espoo, Finland as a case study to compare the IB and FNC instruction of first-year students, the assessment practices of EFL teachers based on the marking of a common essay and a subsequent interview, and quantitative analysis of IB and FNC exam results and essay scores for first-year pre-DP and FNC students in academic year 2019–2020. The mixed methods research (MMR) approach of the thesis is designed to account for the broad set of data (i.e., curriculum, European language policy, academic literature, local needs) that are taken into consideration when developing curriculum at the school level.

The results of this thesis indicate that FNC assessment is more closely aligned with CEFR but that both FNC and IB in Finland, especially pre-DP education, are still lagging behind in implementing the framework developed already two decades ago. At least some of this lag, as indicated in the qualitative data, seem to originate from the continuation of outdated practices of language assessment (esp. in its focus on language mastery rather than action-oriented language use) regardless of changes in CEFR and the Finnish curriculum. In the case of assessment in FNC, the mixed use of continuum criterion-referenced assessment of written production in an exam designed for mastery norm-referenced assessment is an already imperfect combination that is further undermined by lack of transparency around how to apply criterion-referenced assessment to written production as well as how essays marked by teachers have been moderated by sensors.

While the IB is much better at enabling communication between moderators and teachers, EFL instruction in the IB offers a very limited set of communicative language activities, partly due to the broad language profile of students in the international program. For the purposes of students in Finland, however, especially at the school studied, EFL instruction in IB omits an alarmingly wide range of communicative activities in course and exam assessment, an absence that should threaten to invalidate CEFR benchmarking of students graduating from EFL courses, given that most communicative language activities are never taught or assessed.

Many of the issues that arise in this thesis are indicative of systematic issues in the curriculum and examination process, which have a negative washback effect on instruction and assessment at the school level. Nevertheless, some suggestions are made at the end of the thesis for how teachers can mitigate problems with validity in the FNC and IB curricula, though changes in FNC assessment is difficult without transparency from the Finnish Matriculation Board. The results of the thesis also indicate that areas of action-oriented language missing in one curriculum are often well explored in the other, such that serious consideration should be given to closer alignment of pre-DP and first-year FNC English instruction at Finnish schools that offer both programs, an effort towards which this master's thesis is intended to contribute. The comparative weaknesses of EFL instruction and assessment in FNC discussed in this thesis can also be used to inform curricular development of the FNC.

Avainsanat – Nyckelord – Keywords
action-oriented language, written assessment, IB, DP, CEFR, validity

Säilytyspaikka – Förvaringställe – Where deposited
Helsinki University Library

Muita tietoja – Övriga uppgifter – Additional information
All data used for the thesis can be downloaded at https://doi.org/10.5281/zenodo.4460549.

Contents

1 Introduction

Ahead of the new Finnish high school curriculum beginning in 2021 (Opetushallitus 2019), the English teaching staff at a high school in the Helsinki Metropolitan Area, has been exploring the possibility of using the momentum of curricular change to combine the English-language instruction of first-year students enrolled in both the International Baccalaureate (IB) Diploma Program (DP) line and the Finnish national curriculum (FNC) line. This master's thesis seeks to investigate the challenge of combining these two lines of study by investigating how action-oriented language, the foundational second/foreign language teaching methodology in Europe recommended in the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001), is assessed in the written language of first-year high school students at the school.

The problem of first-year instruction in IBDP emerges from its incongruence with FNC. While DP is a two-year program, FNC for high school is a three-year program. Although IB has been taught in Finland since 1990 (albeit in the private setting of the International School of Helsinki) (IBO 2020a), its organization in public schools was not established until 1998 (peruopetuslaki [628/1998], §7; lukiolaki [629/1998], §4), with the principle that instruction of the first year of DP would correspond to first-year instruction in FNC. As such, the 16 schools in Finland that currently offer DP alongside FNC teach a preparatory ad hoc curriculum called pre-DP, the arrangement of which is determined by the DP subjects available at the school, curricular changes in DP and FNC, and developments within each school. That is, the correspondence of pre-DP to the first year of FNC high school is not monitored or defined by the Ministry of Education and Culture, nor was this lack of correspondence identified as a problem in the 2007 report by the committee commissioned by the ministry to investigate the state of IB in Finland (Kansainvälisiä koulutustarpeita käsittelevä työryhmä 2007). This flexibility has enabled pre-DP to serve its preparatory function but has consequently meant that the curriculum of pre-DP, at least at the school studied, has continued to drift further and further away from FNC over the years. The disconnect of pre-DP from FNC is in part necessary, however, given that subject options for DP do not correspond with those for FNC, and even course content can differ radically between curricula within the same subject, not to mention differences in curricular demands unrelated to school subjects.

In addition to the curricular issue posed by pre-DP, the concurrence of IB and FNC at the same school has resulted in social tensions between student cohorts that can emerge from both perceived and real differences in comprehensive school GPA upon admission, ethnicity,

socioeconomic background, mother tongue, English language proficiency, nationality, religion, career paths, and university admissions. Given that exposure to diversity has a well-documented impact on improving cooperation, intercultural competence, critical thinking skills, and multiculturalism, especially among adolescents (e.g., Wells, Fox, and Cordova-Cobo 2016; Loes, Pascarella, and Umbach 2012; Schwarzenthal et al. 2020), the idea of combining the instruction of pre-DP and first-year FNC students at least in English, for which students in both lines at the school studied are required to take three mandatory courses, became all the more appealing as a way of improving dynamics among students across the two curricula. Further benefits of combining first-year English instruction include—and are promoted in FNC and DP curricular material—wider collaboration between teachers of the same subject and across curricula, greater emphasis on the differentiation of instruction, a wider range of cultural and linguistic variety to promote the instruction of EFL/ELF, and the harmonization of instruction and assessment across teachers. The last point may also better enable the balancing of deadlines and overall workload across subjects, one of the student welfare initiatives that the school studied has been attempting to address throughout the 2019–2020 school year, ahead of the curricula changes coming to FNC in 2021.

While these benefits may be both wide and far-reaching in their effect, the prevailing concern among teachers has been the extent to which the instruction of students and their level of preparedness for their respective matriculation exams are compromised. To this end, I have chosen to focus on the one assessment component that is common to the three first-year courses across the two curricula: written production. Given that the development of writing skills is commonly emphasized in both curricula and that action-oriented language is the foundational methodology for both IB and FNC language through its emphasis in CEFR (Opetushallitus 2019, liite 2; UK NARIC 2016, 92–95) and does purportedly guide assessment in both curricula (FMB 2017, 12–14, 16–17; Juurakko-Paavola 2019; IBO 2019b, 35–59), I have chosen action-oriented written production as the primary point of comparison for the two curricula.

Due to the close connection of the research conducted in this thesis with ongoing professional and curricular developments at the school studied, the project takes the form and embodies the philosophy of action research, applying methodological rigor to empirical data to inform changes offered as solutions to problems that arise in the educational setting, developing these changes alongside continual reflection with and collaboration among the teaching staff (Wallace 2006, 12–18). To address the challenge of forming a common approach to the instruction and assessment of writing for first-year IB and FNC students at the school studied, the research seeks to answer the following questions:

1. To what extent do curricula and teaching practices in IBDP and FNC concerning writing reflect the shift from grammar to action-oriented language production, as outlined in CEFR?
2. How valid is assessment of action-oriented written language in both curricula?
3. What are the best practices in both curricula that emphasize the instruction and assessment of action-oriented language?

To answer these questions, I will be exploring the assessment of written production assigned to first-year students in English for school year 2019–2020, an experiment in which teachers will essay the same essay (an IB essay, marked by examiners, and used for rater training), and principles of the instruction and assessment of written language production as identified in the curricula and by the teachers, as well as how they are manifested in the textbooks and other sources used in the classroom and in the matriculation exams of either curricula (Figure 1). While this set of data is

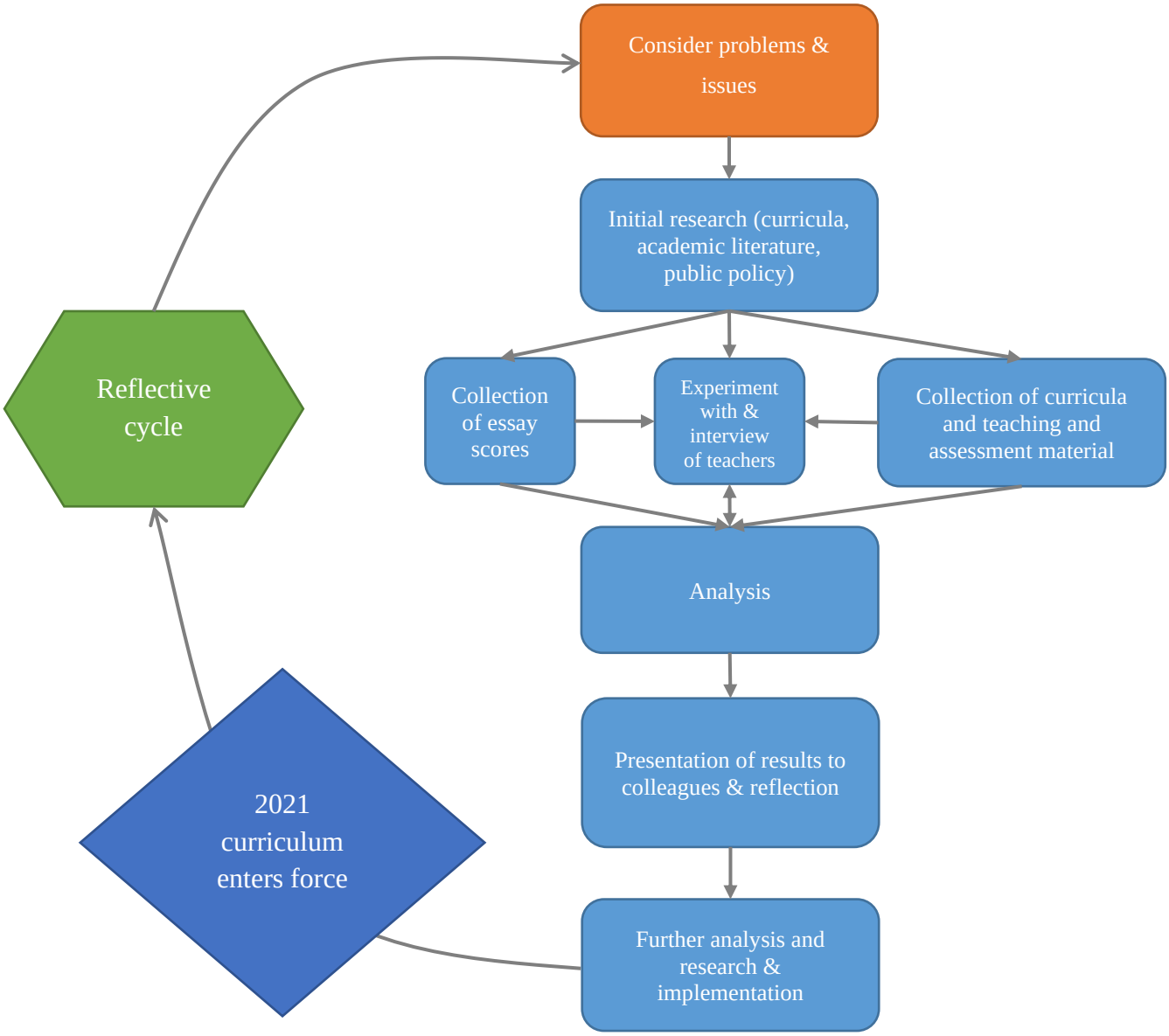

Figure 1: Flow Chart for Action Research (based in Part on Wallace 2006, 14)

both large and broad, their diverse characteristics are the result not of a hopeless endeavor to be comprehensive but of an optimistic attempt to account for the most important and readily accessible factors that apply to the development of a living school curriculum. As indicated in Figure 1, the intention of this master's thesis is to inform professional development based on existing problems. The results of this study will be taken into consideration over the course of school year 2020–2021 to develop the curriculum for first-year English and fully implemented in Fall 2021.

2 Background

The lack of concern by the Finnish education ministry in the uncertain nature of pre-DP in their 2007 report on the state of IB in Finland does not exist in isolation. In fact, there is a lack of academic discussion in Finland on the various issues that pertain to the interaction of IB and FNC within the walls of one school. While student theses on the subject or at least pertaining to IB in Finland abound (e.g., Kolehmainen 2014; Määttä 2014; Hurd 2017; Tamminen 2005; Kovanen 2011; Kainulainen 2006; 2006; Hauta-Aho 2013; Nikku 2019; Karusigarira 2016; Nylund 2017), there are few academic works on the subject (e.g., Junger 1999). This lack of research history may stem from at least three reasons.

First, IB education in Finland, while prestigious (Hansen 2006; Rautio and Niemonen 2020), represents a minority, albeit the largest minority (the other two being the Reifeprüfung Diploma and the European Baccalaureate), of non-FNC high school curricula operating in Finland. As of 2019, the 17 schools offering DP represent only 4.44% of the 383 educational institutions offering high school education in Finland, with only 771 students graduating from the IB in 2019 out of the 37,531 high school graduates in Finland (= 2.05%).

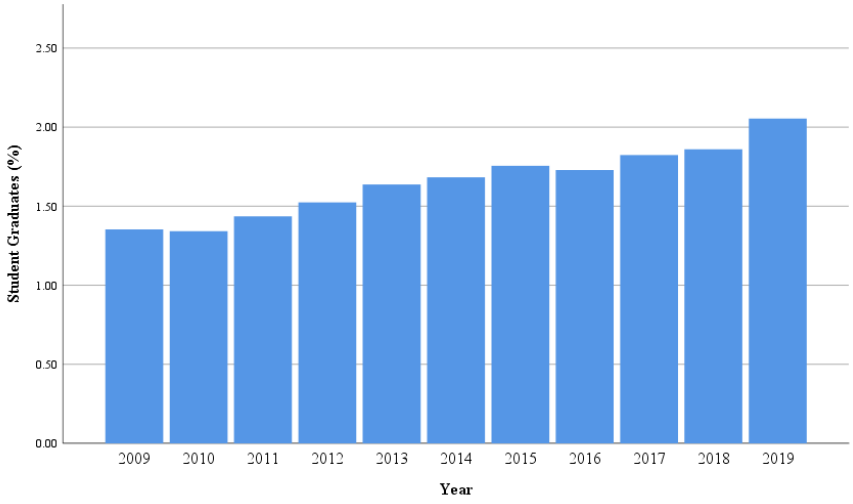

Figure 2: Students Graduating IB as a Percentage of All High School Graduates in Finland (based on data from OSF 2020)

As seen in Figure 2, students graduating IB represent a small but steadily growing population that certainly merits and will likely continue to demand academic attention.

Second, while Finnish curricular material is publicly available on the website of the Ministry of Education, that of the IB is locked off in the closed section (IB Portal) of the IBO website. Nevertheless, access to all curricular information is readily available by inquiry from any IB teacher or coordinator in Finland or elsewhere in the world, meaning that the obstacles for research are equal to that of research in other Finnish educational contexts, where access to school curriculum, practices, scores, etc., would require personal contact with a teacher or personal knowledge of a school.

Third, for professional reasons as well as resource limitations, resolving issues that arise in the interaction of FNC and IB curricula has taken place on an ad hoc basis, varying from school to school and dependent upon regular changes to either curriculum. While IB coordinators in Finland regularly meet one another as a part of institutional practice, subject teachers in the same country are likely to meet one another only during training sessions for revisions in curricula, which occur in the IB subject by subject, and have only recently been possible to have in Finland for subject teachers in Finnish high schools, as the number of schools offering DP in Finland has grown—previously, subject teachers have had to attend curriculum training abroad or online. These occasions for meeting are usually dominated by other concerns (e.g., curricular changes), however, and do not enable a sustained or systematic consideration of IB and FNC interaction. Thus, despite the well intentioned reforms since the early 2000s in Finland towards research-based teacher training (Tirri 2014, 603–5), the sort of academic considerations that inform changes pertaining to IB in Finland tend to be confined within a school's walls rather than evolving as a coherent national approach.

While this master's thesis certainly cannot even dream of being systematic and merely attempts to try to propose reasonable analyses and solutions towards resolving two problems (i.e., cross-curriculum student relations and the growing disconnect of pre-DP from the FNC curriculum) arising at one school, I hope that this thesis can at least prompt further interest in the topic of studying IB in Finland, as FNC and IB curricula and instruction have, as I will try to show, much to learn from one another.

Given the lack of a coherent research history on IB education in Finland, as discussed above, I will focus in the following on addressing existing research traditions pertaining to the learning, instruction, and assessment of written language production in research, especially concerning form-focused instruction research and action-oriented language policy (2.1). I then introduce the two curricula studied (2.2) and key definitions for evaluating assessment (2.3),

concluding this chapter with my hypotheses about the outcome of the study (2.4). In discussing assessment, principles of language learning and instruction necessarily arise as they (1) must be relevant to skills assessed in order for the assessment to be valid and (2) determine what forms of assessment teachers conduct during and at the end of each course.

2.1 Learning, Instruction, and Assessment of EFL/ESL Written Production (CEFR)

While the instruction of L2 languages date back millennia, with a variety of bilingual dictionaries and other documentary texts surviving from antiquity, rigorous research on the effectiveness of different approaches to instructed second language acquisition (SLA) is still nascent. A case in point is that, while competing methods of L2 language instruction flourished throughout the 20th century, heavily impacted by centuries of tradition in the language instruction of Classical Latin (J. C. Richards and Rodgers 2001, 13–16), the first meta-analysis on the effectiveness of different approaches in English L2 instruction was not conducted until the turn of the century (Norris and Ortega 2000; Spada and Lightbown 2013, 320–21). While the foundation for language instruction and assessment in Europe was developed in 1993–1996 and laid down in 2001 with CEFR (Council of Europe 2001, 217), when the effect sizes of L2 instructional methods were only just beginning to be studied (and gaps then and still do persist in the data used for such meta-analyses), CEFR has continued to evolve since its inception, with open-source research on how to improve and implement CEFR in practice publicly available on the Council of Europe website and a companion volume for CEFR published in 2018 that includes new descriptors and scales for language skills (North, Goodier, and Piccardo 2018).

This section begins with a discussion of research on L2 English written production relevant to action-oriented research, focusing on the explicitness and medium of language instruction and assessment and the authenticity of language production (2.1.1) and how these are realized in the 2018 CEFR reference level descriptors (2.1.2). The choice to include the 2018 descriptors necessarily introduces anachronisms to the study when compared to assessment in FNC (2.2) and DP English (2.2.2), given that much of the material in both curricula were developed prior to 2018, but are included anyways given that this study is intended for implementation into evolving teaching practice rather than as a historically focused curricular study on the research leading up to the 2019 IB and the 2021 FNC English curriculum.

2.1.1 Explicitness of Language Instruction (FonF) and CLIL

As the prevailing concern for English teachers at the school studied is the practical syllabus of a combined pre-DP and FNC English course, this sub-section focuses on two primary concerns that apply to the student demographic (see sub-sec 3.2.2, below): explicitness of grammar and pragmatics instruction and the language medium of instruction. Focus on these two aspects of written production often dictate the kind of language tasks and assessments available to teachers.

Research on the effect of the explicitness of the instruction of grammar and pragmatics on SLA has entered into rigorously examination in the past two decades. Norris and Ortega (2000), who conducted the first meta-analysis of research on the topic, observed, among other things, a statistically significant effect size of ½ σ in favor of explicit language instruction for both focus-on-form (FonF) and focus-on-forms (FonFS) instruction across 45 studies (481–82)—with little difference between FonF, the treatment of language as a tool for communication based on the incidental nature of form instruction, and FonFS, the treatment of language as an object to be mastered based on a prescribed syllabus of form instruction (Ellis 2009, 271–306). However, the researchers concede that the ½ σ difference may be unreliable, given that many of the studies used in their meta-analysis were flawed in design: the tests used for treatment effect often design to elicit and even used the terminology of explicit instruction and were highly heterogeneous, with unequal approaches to treatment and a greater number of studies carried out on explicit grammar instruction compared to implicit (70% vs. 30%), likely due to researcher bias.

Shin (2010), revisiting the study, goes a step further than Norris and Ortega's (2000) self-criticism, writing that, despite the seminal impact of their study on SLA research, not only issues with construct validity in the research data used for meta-analysis, but also flaws in Norris and Ortega's own design, oversimplifications made in categorization the studies used, and even their statistical methodology invalidate the study altogether. Most notably, variability in the sample sizes of the studies used for meta-analysis were not adequately accounted for in Norris and Ortega's (2000) study, as well as subsequent studies (e.g., Truscott 2007; Lee and Huang 2008; Spada and Tomita 2010; Goo et al. 2015; Akcin 2019) that have followed their methodology in their decision to calculate effect size using Cohen's *d* (some have also adopted the simplified presentation of FonF and FonFS, in part necessarily due to this issue occurring already in the existing research data), which treats all populations equal regardless of sample size, as opposed to Hedge's adjusted *g* or the Hierarchical Linear Model (HLM), either of the latter of which would better account for variability between and imbalance of sample sizes of studies used for meta-analysis and subsequently their effect sizes in the categories of treatment methods studied. In addition, their decision to focus on only one aspect of FonF instruction meant that different types of FonF and FonFS accounting for

pre-planning of or spontaneity in form-focused instruction, as well as problems in research methodology for eliciting recall that tends to favor FonFS instruction, was not considered (Shin 2010, 25–31; Ellis 2015).

While meta-analyses on the effect of explicit grammar and pragmatics instruction still have severe methodological flaws (in part, besides variability in the sample sizes of constituent studies, because of the difficulty in constituent studies of testing explicit instruction in a way that would be comparable to treatment by implicit instruction), individual studies and theoretical discussions in SLA research do at least suggest, despite existing flaws in the data, that explicit instruction might be more beneficial than implicit, though both approaches demonstrate positive short- and long-term effects compared to control groups (Li 2019, 117–19). The empirical data for this conclusion is based on references to individual studies, however, rather a meta-analysis of relevant studies, while theoretical discussions focus on Schmidt's (1990) Noticing Hypothesis, which postulates that learners must "notice" pragmatic features of language use in order to incorporate such uses into their own interlanguage. Even this tentatively positive conclusion on the explicit instruction of grammar and pragmatics does not account, however, for contextual, age, or attitudinal differences in the learner. Furthermore, comparative studies on inductive versus deductive instruction seem to indicate that FonF instruction, as well as a combination of implicit and explicit instruction, better facilitated the development of especially spontaneous language use, as well as theoretical advantages in terms of learner affect and metacognition, compared to using only FonFS instruction, such that the two approaches to form-based instruction, in both explicit and implicit manifestations, should be utilized (Ellis 2012, 271–336).

Research on Content and Language Integrated Learning (CLIL)—that is, the method of using the target language as the medium of instruction—is, as the *status quaestionis* on the effect of explicit language instruction, in a likewise nascent state in terms of empirical research. Here, longitudinal evidence and theoretical discussions seem to speak in favor of CLIL for the development of both pragmatics and grammar in the target language as offering more opportunities for genuine speech acts by both teacher and student, as well as evidencing a stronger grasp of grammar and pragmatics and broader lexica in written and oral production, especially in higher levels of curricula like upper secondary and higher education (Tateyama 2019; Ryshina-Pankova 2019).

CLIL approaches are not, however, without their limitations. Dalton-Puffer (2011), in her meta-study on the effectiveness of CLIL for SLA, found that, while CLIL learners did consistently outperform their non-CLIL counterparts in oral production, the evidence was more muddled in learners' written production, where the effect of CLIL on morphosyntax and skills beyond the

sentence level (e.g., register, style, genre) are less clear or do not seem to have a discernible effect size compared to traditional forms of foreign language instruction. It should be noted, however, that, like issues with categorizing the learning, instruction, and assessment of FonF in Norris and Oretga (2000), Dalton-Puffer (2011) does not specify here what forms of "traditional" foreign language instruction are compared to CLIL or what differences might exist even within CLIL-based approaches in the studies examined. The meta-analysis is also just a descriptive survey of the research literature rather than a statistically robust comparison of effect sizes produced by CLIL and non-CLIL approaches. Furthermore, the study does not account for the impact of CLIL across levels of proficiency, on which matter Carrió-Pastor and Tamarit Vallés (2015) have suggested, based on a comparison of 50 L1-Spanish higher education students (half enrolled in English or French CLIL program and the other half in a Spanish program), that language acquisition under CLIL environments may be hindered, especially at lower levels of proficiency in the target language.

Another desideratum in CLIL research is the potential cost of CLIL for the development of a learner's L1. As Dalton-Puffer (2011, 189) notes in her study, there is not yet any significant research on this matter. Notably, this line of discussion departs from the debate surrounding the impact of ELF, the primary language of CLIL given that students and teachers are often non-native speakers of the target language (J. C. Richards and Rodgers 2001, 204–20), on mother tongues (e.g., House 2014), since upper secondary CLIL programs like DP come at the cost of upper secondary education in a student's L1, both in terms of exposure to broad academic and social contexts and in terms of translational-grammatical and metalinguistic skills when studying ESL/EFL without reference to the student's L1. In an increasingly globalizing world with limited resources and an ever deepening understanding of what this means for children's language profiles, however, CLIL often emerges as a practical solution to a complex situation, given that increased global mobility means that student's linguistic backgrounds are becoming more and more complicated and diverse (see sub-sec 3.2.2).

In the Finnish contexts of DP as a form of CLIL pedagogy, more research needs still to be done, but two studies conducted for master's theses in Finland seem to suggest positive results for CLIL, with one notably contradicting the assumption that CLIL approaches have a neutral or negative impact on translational-grammatical skills. Karusigarira (2016) found that, in 39 freely written argumentative essays written for the study by 20 second-year FNC students, 10 first-year DP students, and 9 pre-DP students, frequency of errors was consistently higher in the written work of non-CLIL students compared to CLIL students, with roughly half the frequency of errors measured (viz., relating to lexical accuracy, syntax, and cohesion) evidenced in the CLIL corpus compared to the non-CLIL corpus. On the other hand, in her thesis, Hurd (2017) found that, for 57

FNC and 96 IB students from all three years, CLIL students outperformed non-CLIL consistently on 5 English–L1 and 5 L1–English translation questions as a whole, except in the translation of language-specific idioms, where CLIL students performed only slightly better than non-CLIL students on most questions but both groups fared equally poorly (25% accuracy) at translating the phrase "There is no place like home" into their L1 language.

While these two studies offer valuable material for the Finnish context, they do not unfortunately address the question posed here, which is the impact of CLIL on SLA. Neither study looks at the longitudinal effect of CLIL on SLA, as this was not the aim of either study. In fact, at least Karusigarira's (2016) study seems to suggest that, at least at the school studied, differences between CLIL and non-CLIL students arise prior to pre-DP, when students graduating from comprehensive schools self-select as a result of their decision to apply to and their acceptance by schools in Finland offering DP, as well as their presumably higher proficiency in and more positive attitudes towards English. This seems to be the case given that there is only a marginal difference between pre-DP and first-year DP students in Karusigarira's (2016) data when compared to differences between second-year FNC students and pre-DP/DP students as a whole, though her thesis does not provide enough statistical data to re-examine whether or not these differences are statistically significant. Likewise, Hurd (2017) treats performance by students uniformly, regardless of their year of study. Without accounting for the development in student's language ability over time, the methodological framework of both studies would seem to assume that students' language skills do not develop, or do not at least develop significantly compared to those of cross-curriculum peers, over time, an assumption that would need to be checked against their data. Thus, the results from neither thesis can unfortunately be applied to suggesting whether CLIL in the form of DP has a positive impact on SLA in Finland.

The implications for approaches to explicit language instruction and CLIL-based approaches, based on the literature reviewed, suggest that tasks chosen for ESL/EFL instruction of writing skills should be used for explicit, incidental FonF instruction and that CLIL approaches can be beneficial for students at higher levels of proficiency. Proper scaffolding and differentiation, on the other hand, need to be accounted for when considering lower levels of English proficiency, special needs, and, potentially, the development of translation skills. On the research side, more work needs to be done to produce statistically robust meta-analyses on CLIL and explicit language instruction, given the contradictory evidence currently in the field. Especially local data on DP in Finland from academic research would help to measure the effectiveness of DP for L1 and target language acquisition compared to the current approach in FNC and would benefit language instruction in both lines of instruction.

2.1.2 CEFR and the Action-Oriented Approach

The action-oriented approach stems more from the political than the academic realm and is thus a combination of some of the findings by the research surveyed in the preceding sub-section and others that have continued evolve over the past three years since its implementation in CEFR, a framework for language learning, instruction, and assessment published by the Council of Europe in 2001. The connection of CEFR to the institutional project of Europe is undeniable and important to recognize. Despite criticism of the reference-level descriptors (A1, A2, B1, B2, C1, and C2) in SLA research as being an oversimplification of how meaning is successfully communicated in language, its effective presentation and simplicity as well as its implementation into public and university admissions policy and use by standardized testing institutions, the translation of CEFR into over 40 languages (including non-European languages like Esperanto, Mandarin Chinese, and Turkish), its publication in the heyday of the dot com bubble, and its continued incorporation of contemporary linguistic research has meant that CEFR has offered the European project considerable soft power in a post–Cold War globalized world (Figueras 2012; North, Goodier, and Piccardo 2018, 25). The authors thus explicitly emphasize on the first page of the document that its aim is to give "formal recognition to such abilities [as] will help to promote plurilingualism through the learning of a wider variety of European languages" (Council of Europe 2001, 1), a socio-political emphasis reaffirmed in the 2018 companion volume to CEFR (North, Goodier, and Piccardo 2018, 28). The political peacemaking project of plurilingualism will be important especially when considering the plurilingual approach enshrined in the IB curriculum, to which I will return below (see sub-sec 2.2.2).

More relevant for students, workplaces, and institutions for higher education, the CEFR reference descriptors are used to describe the language achievement levels of students graduating from language programs in the IB and FNC, to be used especially for the purposes of university admissions and professional certification (see Table 1). The action-oriented approach and task-based language teaching (TBLT) methodology advocated in CEFR are further, if sometimes implicitly, used to help write the syllabi as well as the criteria developed for the exams in the two curricula (UK NARIC 2016, 15–27; Juurakko-Paavola 2019). This difference in approaches to curriculum writing and examination is perhaps due to the high practical pressure from university admissions offices for language proficiency scores compared to the low legal pressure of CEFR as a piece of legally non-binding, albeit highly influential, public policy.

As mentioned earlier, the action-oriented approach is a combination of different ideas borrowed from SLA research. Its emphasis is on the varying use of language "in real life situations," by "seeing learners as language users and social agents, and thus seeing language as a

Curriculum	*DP*				*FNC*
Course option[1]	English A: Language and Literature (SL)	English A: Language and Literature (HL)	English A: Literature (SL)	English A: Literature (HL)	N/A
Based on overall grade	Yes	Yes	Yes	Yes	No
B1.2[2]		4	4	4	English, B1-level
B2.1	5	4/5	4/5	4/5	English, A-level
B2.2	6	5/6	5/6	5/6	
C1	7	6/7	6/7	6/7	
C2		7	7	7	

[1]On the course options for English in DP, see sub-sec 2.2.2.
[2]Lower reference levels are not given below a certain score for DP Language A courses, since such scores may be indicative of inadequate subject knowledge, though higher scores are suggested not to be influenced by the same factor, since higher scores were found to necessarily demand a high degree of reading comprehension and production skills (UK NARIC 2016, 28–29).

Table 1: Comparison of CEFR Reference Levels in DP and FNC English Course Options (UK NARIC 2016, 95, 110, 131, 146; Opetushallitus 2019, 177)

vehicle for communication rather than as a subject to study," though this emphasis is made without a specific prescription of how much—or, indeed, even if—grammar and literature should be an explicit part of language learning, instruction, and assessment (North, Goodier, and Piccardo 2018, 27). The emphasis on language as a "vehicle for communication" seems to be modeled especially on Long's 1996 hypothesis that language acquisition is most effectively increased when learners are forced to "negotiate for meaning," or to seek to understanding meaning when communication breaks down, as well as his Long's (1991) influential distinction of FonF from the traditional FonFS model, the former of which specifically emphasizes drawing attention to form (= secondary) only while learners are communicating (= primary), as opposed to the traditional FonFS model, which treats linguistic forms as the primary goal of language learning and their realization in varying social situations as a secondary goal, if at all. As explored in the previous sub-section, current data are still inadequate even to suggest that one approach has a larger effect size than the other, but the innovative emphasis and appeal of CEFR to the nascent field of Form-Focused Instruction in 2001 has made its emphasis on the social context of language use a defining feature of its language teaching philosophy.

To achieve the instruction of language in varying social context, CEFR recommends a primarily TBLT-based approach on the merits of its cognitive, affective, and linguistic benefits (Council of Europe 2001, 157–67). That is, the small scale of individual language tasks enables a

manageable cognitive load for students, familiarity with certain task types can help promote motivation, risk-taking, cooperation, and a sense of growth while introducing new sociocultural knowledge, and the range of possible tasks enables broad development of linguistic skills, attention to interlanguage, and possibilities for differentiation that can also be assessed for achievement. The positive evaluation of TBLT for L2 acquisition in CEFR has been echoed in empirical and theoretical studies in SLA research, especially as "authentic" language material becomes increasingly available and has even developed exclusively online, and the need to cultivate language alongside ICT skills continues to grows today (J. C. Richards and Rodgers 2001, 223–43; González-Lloret 2019).

In terms of written production, which is distinguished from written interaction and mediation in CEFR, the two main categories described are (1) creative writing and (2) written reports and essays, which are complemented by three production strategies: (1) planning, (2) compensating, and (3) monitoring and repairing (Council of Europe 2001, 61–65; North, Goodier, and Piccardo 2018, 75–80). The writing task types categorized in CEFR as communicative language activities and strategies are given in Table 2—note that what CEFR terms "communicative language activities" are referred to as task types in this thesis. Comparing the mean/median score of long-form studies in English in DP and FNC for Spring 2019 with their CEFR benchmarking (see Tables 1, 14), both curricula seem to target a completion of curricula at the B2.1 level (at the 94th/68th percentile in the case of DP scores), the descriptors for which in communicative language strategies (i.e., production, interaction, and mediation)[1] and competences (i.e., linguistic, sociolinguistic, and pragmatic) are given in Appendix A.

Communicative Language Activities and Strategies	**Text type**	**Brief description**
Production (Sustained language writing without immediate reaction from recipients)	Creative Writing	"personal, imaginative expression in a variety of text types"
	Written reports & essays	"from short reports and posters, to complex texts which present a case, or give critical appreciation of proposals or literary works"
Interaction (Writing in which communication is co-constructed by two or more participants)	Correspondence	"from simple, personal messages, to in-depth, personal and professional correspondence"
	Notes, messages & forms	"a range of transactional interactive writing"

[1] Reception, the fourth type of communicative language activity and strategy, is not included here, since it does not involve writing.

	Online conversation & discussion	"group interaction online that are almost impossible to capture in traditional competence scales"
	Goal-oriented online transaction & collaboration	"potentially collaborative nature of online interaction and transactions that have specific goals… a rigid separation between written and oral does not really apply to online transactions"
Mediation (Written communication across texts, languages, dialects, cultures, or other contexts)	Relaying specific information in writing	"the way some particular piece(s) of information of immediate relevance is extracted from the target text and relayed to someone else"
	Explaining data in writing	"the transformation into a verbal text of information found in diagrams, charts, figures and other images"
	Processing text in writing	"understanding the information and/or arguments included in the source text and then transferring these to another text, usually in a more condensed form, in a way that is appropriate to the context of situation"
	Translating a written text in writing	"a largely informal activity that is by no means uncommon in everyday personal and professional life"
	Note taking	"the ability to listen and write coherent notes"
	Expressing a personal response to creative texts	"expression of the effect a work of literature has on the user/learner as an individual"
	Analysis and criticism of creative texts	"more formal, intellectual reactions [than expressing a person response to creative texts]"

Table 2: Communicative Language Task Types Relevant to Writing in CEFR (2018)

The action-oriented approach is notably evident in the 2018 CEFR reference descriptors for communicative language strategies in their emphasis on different social contexts and the need to negotiate for meaning. Variety of contexts, for instance, can be seen in the inclusion of creative writing (an oft-neglected aspect of English language instruction), types of letters and e-mails, degrees of formality, different online platforms, a consideration of the audience of a piece of communication. Likewise, the need to negotiate for meaning is evident in the need for B2.1 language users to have a relatively strong ability to understand their own mistakes and those of

others in communication, correct their own mistakes, seek clarification or elaboration in different media, and use circumlocutions where language users are unable to express themselves.

The inclusion of linguistic, sociolinguistic, and pragmatic considerations in the descriptors for communicative language competences further affirms how CEFR has been developed alongside contemporary research in applied linguistics, applied psychology, and socio-political studies. As emphasized by the modifier "communicative" in this category, the focus in each criterion is on the extent to which a certain linguistic competence enables its user to communicate, rather than to demonstrate mastery of a given form. This emphasis is clearly an FonF approach, as is evidence in the description of communicative language competences as a whole: "these aspects, or parameters of description, are always intertwined in any language use; they are not separate 'components' and cannot be isolated from each" (North, Goodier, and Piccardo 2018, 130). Thus, even in these criteria, the descriptors refer to the context, style, and audience of writing. The formulation of the criteria are also done in a way that focus on what learners are able to do achieve than what they fail to do, in line with how studies in positive psychology have promoted the benefits of a growth mindset (Dweck 2006). In addition, from a socio-political perspective, the criteria blur the distinction between multilingualism and multidialectalism, emphasizing their equality, importance, and troubled history in language learning.

Finally, the reference descriptors for plurilingual and pluricultural competences attempt to capture the complexity of language profiles in a globalized world. Strategies for using one's various languages to facilitate communication or language learning, for instance, are taken into consideration, as well as one's ability to adapt not only to different social context but also different and new cultural contexts and cultivating a familiarity of how "otherness" might arise in language.

Due to the wide audience of CEFR, there are no specific guidelines for which assessment should be applied to the varying strategies and competences. The document instead offers a broad consideration of 26 different approaches to assessment across 13 different categories, describing and briefly evaluating each approach (Council of Europe 2001, 183–92). Specific, simplified criteria have also been provided as diagnostics for holistic, proficiency, criterion-referenced self-assessment and written assessment (North, Goodier, and Piccardo 2018, 167–70, 173–74). Both FNC and DP English syllabi prescribe a range of approaches assessments, but the most relevant one for this thesis are those relevant to learning, instruction, and assessment of the written production in the two curricula, discussed in the following sections.

2.2 Curricula

The two syllabi examined in this study are the 2016 and 2021 English A syllabi for the FNC curriculum (0) and the 2013 and 2019 English A: Language and Literature syllabi for the IB curriculum (2.2.2). The year corresponding to the syllabus is given here denoting the first year of instruction rather than the first year of examination. The newest syllabi in both curricula are used whenever possible to evaluate their implementation of action-oriented language instruction and assessment, but there are some cases of analysis in which the older syllabus is more relevant, summarized in Table 3. When comparing exams, this thesis looks only at the recently digitalized matriculation exam for second national and foreign languages, A syllabus (FMB 2017). The exam in the IB, on the other hand, corresponds to the syllabus.

	FNC		**IB**	
	2016	2021	2013	2019
Most recent exam in both curricula (May 2019)	✓		✓	
Course instruction and assessment for school year 2019–2020	✓			✓
Curricular content and goals		✓		✓
Interviews with teachers	✓	✓	✓	✓

Table 3: Syllabi of FNC and IB Curricula Used for Comparison

2.2.1 FNC, English as a Foreign Language

The large part of the instruction and assessment of FNC students is separately treated by the national curriculum, set by the Ministry of Education (*opetushalitus*) (lukiolaki [714/2018], §§10–18), and the Finnish Matriculation Board (FMB, *ylioppilastutkintolautakunta*), a separate body appointed by the Ministry of Education in three-year terms to oversee the matriculation exams (laki ylioppilastutkinnosta [502/2019]). Separate curricula expanding on the national curriculum is then developed formally by municipality and schools with special mandates, based on needs especially arising from differing demographics and histories, and informally by school. The current FNC syllabus for English is the 2016 curriculum, to be renewed in 2021, with explicit recommendations in both to use CEFR to support self-, peer-, and teacher-assessment (Opetushallitus 2015, 108; 2019, 178). Its description also reflects the impact of CEFR in its explicit description of the foreign language syllabus as developing communicative language strategies and competences (with examples of reception, production, interaction, and mediation), as well as plurilingual and pluricultural competences (Opetushallitus 2019, 174).

Consistent with other countries that have high-stakes matriculation exams (especially when used for university admissions) and are subsequently impacted heavily by the washback effect (Ferman 2004), the pedagogical goals of curricula are often sidelined by textbooks, teachers, and

students, for whom the preference is often to achieve good end results, as measured by the matriculation exam.[2] In this sense—for both FNC and DP—curriculum, exam, and instruction are linked together inextricably.

In the 2016 English syllabus, teaching is divided in the national curriculum across six mandatory and two optional courses (as well as an optional matriculation exam preparation course not specified in the curriculum but effectively offered in schools for each subject), with one course lasting one-fifth of a school term. Across the roughly two-and-a-half years of high school (= 12 teaching periods), English instruction thus accounts for 50–67% of a student's high school education in FNC. In their first year of instruction, students at the school studied are required to the first three courses of English (see Table 4), given that they have consistently studied English as their first foreign language since comprehensive school. In the table, the literary and non-literary reading material are also given not as examples of reception task types but primarily as source texts for mediation and potentially also as models for production.

		FNC	*Pre-DP*
Course 1	**Name**	**The English Language and My World**	**Language: Readers, Writers, and Texts**
	Literary texts	N/A	Poetry, song lyric, spoken word
	Non-literary texts	TED talk, travelogue, online conversation	Advertisement
	Written assessment	Production (essay, argumentative or reflective)	Production/Mediation (essay, guided textual analysis)
	Other written tasks	E-mail	Poetry, learner portfolio
	Textbook	*New Profiles 1* (Lindroth, Hannuksela, and Rosenback 2016)	N/A
Course 2	**Name**	**The Human in Networks**	**Mermaids in Texts**
	Literary texts	N/A	Poetry, scripture, short story, song lyric
	Non-literary texts	News article, TED talk, website	Academic writing, blog post, comic, film, internet memes, musical theater, music video, news article, online review, painting, parody, photography, reaction gifs, speech, statue, travel guide
	Written assessment	Production (essay, argumentative or reflective)	Production/Mediation (essay, comparative literary/media analysis)
	Other written tasks		Learner portfolio
	Textbook	*New Profiles 2* (Andtfolk, Hannuksela, and Lindroth 2016)	N/A
Course 3	**Name**	**Cultural Phenomena**	**Culture: Time and Space**
	Literary Texts	Hymn, novel	Short story
	Non-Literary Text	Documentary short, graffiti, new article	Academic writing, advertisement, film, institutional rules, music video, manifesto, news article, opinion piece,

[2] Although I was not able to find research on the washback effect in the context of Finnish high schools, except in the context of the Ingrian language (Martikainen 2020), I can at least point to the washback effect at the school studied in this thesis based on interviews with the teachers and a survey of the assessment and instructional material (see sub-sec 4.2.2).

		oral narrative, parody, photography, tweet
Written assessment[1]	• Production (essay, reflective) • Production/Mediation (essay, reflection on literature) • Production/Mediation (creative, fictional letter) • Production (creative, ghost story) • Interaction (letter to newspaper editor) • Interaction (letter, reflective)	Production/Mediation (essay, comparative literary/media analysis)
Other written tasks		Learner portfolio
Textbook	*New Profiles 3* (Rosenback et al. 2017)	N/A

[1]Bullet points indicate that the options offered for written assessment included several different task types. The category of production, interaction, and mediation are based on the categorization of communicative learning tasks in CEFR.

Table 4: Course Distribution, Reading Material, and Written Assessment for First-Year Students at the School Studied (data gathered from unpublished documents stored in the city-purchased G Suite as part of the school curriculum)

Preference for English as the primary foreign language (as opposed to studies in "Mother Tongue and Literature" and the "Second National Language") to be studied is clear from how the syllabus for how foreign languages is written, with the possible A-languages to be studied in both the 2016 and the 2021 curricula given in the following order: English, (other) foreign languages, Asian and African languages, and Sámi. While this categorization is well intentioned, with different socio-cultural consideration given to the four categories of A-languages and how they should be taught, the hierarchical nature of the list and the assumption of normality in the "(other) foreign languages" category over against Asian and African languages or Sámi are problems worth investigating but beyond the scope of this thesis. I would at least point out that the treatment of foreign languages in FNC differs drastically the more democratic and uniform vision of CEFR. It is also worth noting that the format of and instruction of the digital exam for all foreign language A syllabi, including the second national language syllabus, are the same (FMB 2017).

Written tasks types for foreign languages in the matriculation exam are not prescribed in law, nor are they specified in material distributed by the FMB, except that students are to write one text of 700–1,300 characters in length based on one of several questions (FMB 2017, 12). Since the digital exams for the English A syllabus in FNC started in Fall 2018, this gives a feasible time constraint (i.e., 2018–2020) to the data set; the list of text types for exams in FNC and DP is given in Table 5, based on the three categories for written task types given in CEFR (see Table 2).

		FNC[2] *(N = 25*[4]*)*	*DP*[3] *(N = 32)*
Production[1]	Creative Writing		
	Written reports	✓	
	Essays	✓	✓
Interaction	Correspondence	✓	
	Notes, messages & forms		

Mediation	Online conversation & discussion	✓	
	Goal-oriented online transaction & collaboration		
	Relaying specific information in writing	✓	✓
	Explaining data in writing	✓	✓
	Processing text in writing	✓	✓
	Translating a written text in writing	✓	
	Note taking		
	Expressing a personal response to creative texts	✓	
	Analysis and criticism of creative texts		✓

[1]FNC written texts included prompts that students to write speeches that could be categorized could be categorized as spoken production. These are not included here, since the premise is that students would only write the text and not deliver it. In addition, while written reports and essays are grouped together in CEFR, they are included here separately to further distinguish FNC from DP data.
[2]Data were retrieved from Yle Abitreenit (https://yle.fi/aihe/artikkeli/2015/12/15/yo-kokeet-englanti).
[3]Data were retrieved from the online IB Programme Resource Centre (https://internationalbaccalaureate.force.com/ibportal/apex/ibportallanding). Note that no Spring 2020 exam was given in DP due to coronavirus pandemic (IBO 2020c). The exams used for this data set were English A: Language and Literature (HL).
[4]This number reflects only the number of questions in the written section of the exam (also in the case of the DP figure). Other questions in the exam were used as examples of "Mediation: Translating a written text in writing" but not counted toward this figure. While some written sections of the FNC exams resembled "Mediation: Note taking," the source texts for these tasks differed markedly from the source text types described in CEFR, and the task itself also differed in substance.

Table 5: Written Task Types in DP and FNC Exams, Based on CEFR (2018) Categories, 2018–2020

2.2.2 DP, English A: Language and Literature & pre-DP English

The language A syllabus in DP, in contrast to its presentation in FNC, is presented as a single syllabus for all 55 languages examined as a language A in Spring 2019. This uniformity is reflective of the role language has as a peacemaking project, in common between CEFR and the IB. One of the foundational documents for the IB educational philosophy, was an essay entitled "Educational Techniques for Peace: Do They Exist?" (1948), commissioned by UNESCO and written by Marie-Thérèse Maurette (1890–1989), a French educator at the Geneva International School who had gained international attention for her peace-driven pedagogical philosophy. In the essay, Maurette outlines four conditions that she deemed essential for children to "become members of the human race as a whole, and not merely members of separate nations":

1. avoiding nationalist biases in the youth by mixing children of different nationalities together;
2. incorporating as whole of a global perspective as possible into early childhood education;
3. weakening notions of "foreignness" through second language education by native speakers;
4. fostering community dynamics and social interdependence within the school. (Maurette 1948, 17–18)

The third point, though controversial now for its promotion of language education through native speakers, is notable for its emphasis on the explicit aim to reduce otherness through language

education that would later be echoed in CEFR (Council of Europe 2001, 1; North, Goodier, and Piccardo 2018, 22). This point again stands in stark contrast to the FNC approach to language A syllabi, which envisions English, (European) languages, Asian and African languages, and Sámi as essentially different language groups. The first point in Maurette's list is also interesting, as it was a fundamental assumption at the school studied that tensions between IB and FNC students arose because the school (and other schools in Finland that offer DP alongside FNC) had failed to mix the groups sufficiently, leading to racially and religiously insensitive attitudes in both groups of students. This assumption was a point of departure for this action research.

Following Maurette's thinking, the IB curriculum is explicitly aimed at peacemaking. The mission statement and learner profile, included in every DP syllabus, states that the aim of the IB is "to create a better and more peaceful world through intercultural understanding and respect" and that their students should be "internationally minded people who, recognizing their common humanity and shared guardianship of the planet, help to create a better and more peaceful world" (IBO 2019b). It is perhaps in part because of this shared socio-cultural emphasis on peace through plurilingualism and pluriculturalism that the DP language and literature syllabus focuses on the study of "communicative acts across literary form and textual type" (IBO 2019b, 7), though this description does not, admittedly, explicitly mention language, dialect, or culture as intermediaries for communicative acts.

There are, at any rate, logistical reasons for this emphasis on communicative acts. In DP, there is no separate mother tongue studies compared to other advanced studies of languages. Unlike FNC, DP is an international curriculum, meaning that students' experience with different languages and cultures will vary much more widely. Thus, student language profiles in DP are intentionally vague. The B syllabus for DP language studies is described as being "designed for students with some previous experience of the target language" (IBO [2018] 2019, 6), while the A syllabus, the only one currently offered for DP English at the school studied, is described as being "for students from a wide variety of linguistic and cultural backgrounds, who have experience of using the language of the course in an educational context" (IBO 2019b, 6). Language placement decision is based on these descriptions at the discretion of each school's administration, including the DP coordinator. Indeed, as indicated in Table 9, 18% of the pre-DP students (vs. 0% in FNC) for school year 2019–2020 had English registered as their mother tongue and will be placed alongside students with English as their L2 when entering DP for school year 2020–2021. Thus, the DP Language A syllabus is neither just a foreign language syllabus nor a mother tongue syllabus but a combination of the two, though to varying degrees.

Students enrolled in DP have three course options for their Language A study: (1) Literature, (2) Language and Literature, and (3) Literature and Performance. Of these three, only the Language and Literature syllabus approaches language broadly, with an emphasis on the study of communicative acts (IBO 2019b, 6–7), best representing the full range of strategies and competences envision in CEFR. Furthermore, students can opt for standard-level (SL), constituting 150 teaching hours, or higher-level (HL) studies, constituting 240 teaching hours, in their Language A study. To study in DP, students are required to enroll in three HL subjects and three SL subjects (with exception made for students opting for four HL and two SL subjects) across six academic areas, languages A and B being categorized separately. At the school studied, SL and HL courses run for the entirety of the high school, or 12.5 teaching periods (half a period longer than in FNC), with different weekly course-loads assigned to HL and SL subjects. As shown in Table 14, the global tendency, which is reflected at the school, is for students to study English as an HL subject.

The washback effect discussed in the context of FNC is not entirely absent from DP instruction. However, the washback effect may or may not be mitigated by the DP subjects being taught for the entirety of a student's IB education, the design of the curriculum, and the final grade being only partially determined by the final exam. While the second point will be considered from the interview and thus discussed later, the third point can be briefly summarized in Table 6. As indicated there, the final exam in DP English A studies accounts for only 60% (HL)/70% (SL), thus representing slightly a slightly lower cognitive load compared to the matriculation exam in FNC, which is determined entirely on the exam date, regardless of grading across school courses.

		HL	*SL*
Final Exam	Paper 1 (Guided textual analysis)	35%	35%
	Paper 2 (Comparative Essay)	25%	35%
Examiner-assessed (written during school year)	HL essay (Analytical essay on literary and/or non-literary texts)	20%	N/A
Teacher-assessed (moderated by the IB)	Individual Oral (Comparative commentary on non-literary and literary texts)	20%	30%

Table 6: Assessment Outline for HL and SL Studies in DP English A: Language and Literature (IBO 2019b, 33–34)

Language education in DP is effectively built on CLIL pedagogy, and, while the official languages of DP are English, Spanish, and French, all schools offering DP in Finland operate only in English. Thus, DP subjects, including languages A and B, may be taught by non-native speakers of English, such that ELF emerges as the primary language of DP.

At the school studied, pre-DP English is modeled on the current FNC English A and the IB English A: Language and Literature syllabi, regardless of differing demands between HL and SL (Table 4) or students choosing the English A: Literature option in their DP studies (see sub-sec 2.2.2). Because HL studies in the English A: Language and Literature syllabus are preferred by

students to SL studies, though not to the same extent as long-form English matriculation exams (Table 14), direct comparisons between the two curricula/exams in this thesis are made between HL English A: Language and Literature (DP) and the Foreign Language A: English (FNC), with the matriculation exam in long-form English A used for comparison to the DP exam.

2.3 Assessment of Written Production

As noted earlier, approaches to assessment vary widely between FNC and DP, and there is no prescribed form of assessment outside of certain principles for summative and performance assessment in FNC and DP English instruction, while no single form of assessment is recommended above others in CEFR, though the tendency is to recommend all forms of assessment in moderation. This section will define only the most relevant forms of assessment based on their description in CEFR (Council of Europe 2001, 183–92). The following section discusses only the forms of assessment most relevant to this thesis, especially pertaining to the course content of first-year English instruction at the school studied, the interview data, and exam score comparisons, before going in to a discussion of assessment validity (2.3.1) and written assessment in FNC and IB (2.3.2). To facilitate the readability of this section, each form of assessment is italicized when introduced alongside the definition used in this thesis. I have chosen to group each form of assessment by common theme, though they can overlap across and even within categories.

Assessment can, first, vary in terms of the object measured. Assessment can be made internally to the objectives of a syllabus, *achievement assessment*, or in relation to the real world, *proficiency assessment*. Marks awarded within each course are generally almost always achievement assessments, indicative of a student's progress in their subject studies, while marks awarded for externally assessed components, including a student's matriculation exams, are almost always proficiency assessments, normalized for use in institutional and professional settings like university admissions or professional qualifications, resulting in a high consequential value ascribed to such forms of proficiency assessment, or *high-stakes assessment* (Shaw and Weir 2007, 226). *Direct assessment* measures the strategies and competences while the candidate in engaged in those activities, while *indirect assessment* relies on the results of various forms of strategies and competences that have come together as a final result, especially in the form of an exam. *Holistic assessment* synthesizes several aspects as a global judgement, while *analytical assessment* distinguishes between several, discrete aspects of the object measured.

Assessment can likewise vary based on the subject applying assessment. *Subjective assessment* prioritizes decisions being by an assessor, while *objective assessment* prioritizes made

in test design, usually in the form of multiple or fill-in-the-blank questions. *Assessment by others* can be done by peers, teachers, or examiners, while *self-assessment* is done by the learners themselves. The latter especially has been associated with positive affective effects, especially in terms of student motivation, mindset, metacognition, and self-orientation (Council of Europe 2001, 191–92; Dweck 2006, 28–29; McMillan and Hearn 2008).

Assessment can vary based on how degrees of difference are referenced. Assessment based on criteria can refer to performance within a population, *norm-referenced (NR) assessment*, or can be based on fixed criteria regardless of performance within a population, *criterion-reference (CR) assessment*. Within CR assessment, examiners can develop criteria that determine the minimum level of achievement or proficiency needed, the *mastery CR approach*, or that indicate a range of achievement of proficiency, the *continuum CR approach*.

Assessment across a length of time can differ in several ways. *Continuous assessment* describes the use of material produced throughout a course or diploma, while *fixed-point assessment* describes decisions made based on exams from a single day. On the other hand, assessment can be given as an ongoing process of studies, *formative assessment*, and as summation at the end of a task or the course of a school term or diploma, *summative assessment*.

2.3.1 Validity of Assessment

Validity of assessment describes the proximity of the gap between the assessment and what is assessed (= confidence), as well as the consistency with which that assessment is applied across time, space, and population (= precision), within the context of the administration of the assessment (Weir 2005, 11–16). As with assessment, there is a wide range of approaches to validity, considering *a priori* and *a posterori* factors of the test, or factors that can be considered before and after the event of the test (Weir 2005, 47). Table 7 thus offers a survey of the aspects of validity relevant to writing drawn from their presentation in Shaw and Weir 2007. While the monograph focuses on validity of assessment within the institutional setting of the Cambridge ESOL environment and tests like it (e.g., TOEFL, IELTS), it serves as a broad secondary source on written assessment that incorporates evidence and research from the field, and the institutional focus, rather than detracting from the usefulness of the monograph, helps serve as a model for how to think about the DP or FNC exam environment. The categories of validity are given alongside their definition and parameters, with the latter based largely on Shaw and Weir 2007 but with some adjustments made to match the terminology used in this thesis.

	Category	Definition	Parameters
***a priori* validity evidence**	test-taker characteristics	"personal characteristics of the individual test taker" (17)	**physical/physiological** (age, gender, special needs), **psychological** (special needs, motivation, multiplicity of tasks and types), **experiential** (familiarity with exam format and setting)
	cognitive validity	"how closely [a writing task] represents the cognitive processing involved in writing context beyond the test itself" (34)	**macro-planning** (determination of what ideas, genre, etc., are needed for task completion), **organization** (order, focus, and cohesion of ideas), **micro-planning** (sentence- and paragraph-level planning), **translation** (conversion of abstract content to linguistic expression), and **finesse** (mechanical accuracy, style, register) demanded and **revision** (sentence-, paragraph-, and text-level corrections) allowed
	context validity	"the linguistic and content demands that must be met for successful task realization and to features of the task setting that serve to describe the performance required" (63)	**task** (task type & format, purpose, knowledge of criteria, text length, time constraints, writer–reader relationship), **administration** (physical conditions, uniformity of administration, security), **linguistic demand** (lexical resources, structural resources, discourse mode, functional resources, content knowledge)
***a posteriori* validity evidence**	scoring validity	"the extent to which test scores are based on appropriate criteria, exhibit consensual agreement in marking, are free from measurement error, stable over time, consistent in terms of content sampling and engender confidence as reliable decision-making indicators" (143)	**criteria/rating scales** (holistic vs. analytical assessment), **rater characteristics** (physical/physiological, psychological, experiential), **rating process** (rater expectations and reliability), **rating conditions** (venue of examination or score awarding, handwriting vs. word-processing, time, scaffolding), **rater training** (for consistency as well as minimization of biases concerning task types), **post-exam adjustment** (NR vs. CR assessment), **grading and awarding** (distribution and publication of results and certificates)
	consequential validity	the extent to which a test changes the behavior or attitudes of learners, instructors, and institutions (218–20)	**washback on individuals** (specific to the classroom, the positive effects of which are contingent upon proper training), **impact on institutions and society** (socio-political and psychological empowerment or disempowerment), **avoidance of test bias** (recalibration of test based on relationship between exam results and test-taker characteristics)
	criterion-related validity	"the extent to which test scores correlate with a suitable external criterion of performance with established properties" (229)	**cross-test comparability** (comparability across tests with similar objects of assessment), **within-test comparability** (comparability across time, space, and population), **comparability against external standards** (standardization with CEFR and evolving research)

Table 7: Validity in the Assessment of L2 Writing (Shaw and Weir 2007, with some adjustments)

In each category except for consequential validity, degree of validity is inversely correlated with effect size on exam scores, using factors indicated in each parameter. It should be noted that empirical research has yet to be conducted on every aspect outlined in Shaw and Weir 2007, but there is broad consensus in the field that the factors covered should be considered when evaluating the validity of assessment. In the case of consequential validity, Shaw and Weir 2007 tentatively suggest a positive correlation of consequential validity with its parameters. It should be noted, however, that the authors are closely associated with the Cambridge ESOL exams, and their conflict of interest in making this evaluation is not properly acknowledged.

In my view, the argument that washback can be positive assumes that there are short-comings in the validity of curriculum or the realization of the curriculum in the cultivation of communicative language strategies and competences and plurilingual and pluricultural competences that do not exist high-stakes exams. As a teacher, it is perhaps unsurprising that I would argue that washback tends not to be positive, because such strategies and competences can be better gauged longitudinally (= classroom) than in a single test environment (= high-stakes exam). However, I would concede that differences necessarily exist in the realization of curricula, including the effects of hidden curricula, and that, here at least, high-stakes exams can act as a beneficial corrective. Thus, further sub-parameters, as free of research bias as possible, would need to be developed to better appreciate the negative and positive effects of backwash.

Impact of high-stakes examination is, however, determined almost entirely by one's socio-political viewpoint. In Finland, where—despite relatively high income equality, with a Gini coefficient of 27.4, within the top 15 countries in the world (World Bank 2020)—social capital is still strongly correlated with educational background (Rinta-Kiikka, Yrjölä, and Alho 2018). The high school diploma is regarded as an object of prestige, with the last day of school before the high school matriculation exams widely celebrated and televised (*penkinpainajaiset*) and the high school graduating cap (*ylioppilaslakki*) worn proudly on May Day. Ironically, while the latter holiday is purportedly a celebration of the worker, and graduates of vocational school also don their hats (*ammattilakki*) proudly on the day, the festivities culminate in the capital with the televised donning of the *ylioppilaslakki* on Manta, a prominent statue overlooking a market-square traditionally used by fishermen. Similarly, differences in the test-taker characteristics can have an outsize impact on test results, where discrimination, whether related to socio-economic status, gender, sexuality, religion, or immigrant background is entrenched in society, as is the case in Finland regarding "large performance gaps" related at least to gender, immigrant background, and native language (OECD 2015). Even with the best statistical modeling, examination institutions would be unable to account for all these differences, despite the high stakes of their exams, such that they can, at best, reproduce the discrimination inherent in educational systems and, at worst, exacerbate these issues. In the school studied, perceived differences about the difficulty of the exams have also resulted in different attitudes about either curriculum as a whole, especially regarding perceived English proficiency among IB students. Care should be taken, therefore, when considering the relationship between impact of high-stakes examination and consequential validity.

2.3.2 Assessment in FNC and IB

Course assessment in FNC and pre-DP are functionally the same, based on the 4–10 scale prescribed in FNC (Opetushallitus 2019, 47), which is defined in valtioneuvoston asetus lukiokoulutuksesta (810/2018, §17) as follow: 4 (fail), 5 (adequate), 6 (moderate), 7 (satisfactory), 8 (good), 9 (commendable), 10 (excellent knowledge and skills demonstrated). As there is no prescribed percentile grade boundary, grade boundaries are developed at schools and by teachers and may take the form of either NR or CR assessment.

Proficiency assessment, as indicated in Table 1, is aligned with CEFR in both exams. The DP exams use analytical, weak continuum CR assessment, while FNC exams results are processed as norm-referenced analytical assessment, the results of which are used to determine mastery CR assessment. That is, in DP, the resulting seven-point grade calculated from exams and other assessment components (Table 6) are compared to set grade descriptors, developed for each of the six academic areas of study (IBO 2017), with a continuous proficiency level that can range of B1.2–C2. These results are finally reworked comparing these grade boundaries to those of previous years (IBO 2018, 77). FNC exams, on the other hand, norm-reference the grades of first-time test-taking candidates based on a set distribution for their seven-point scale: 1 (5%), 2 (15%), 3 (20%), 4 (20%), 5 (20%), 6 (15%), 7 (5%). These grade boundaries are then set for all candidates, with a passing grade of 2 or higher used to determine the fixed proficiency level of B2.1.

Examination processes also differ between the curricula. In DP, exams are marked anonymously and externally by two examiners, whose results must agree within tolerance (usually 10% of the overall score, though this depends on the decision made by the chief examiner), or else the exam is sent to a principal examiner for arbitration (IBO 2018, 115–16). Thus, not all teachers in DP subjects are, or indeed can be, examiners in the subject they teach. In FNC, on the other hand, exams are preliminarily marked non-anonymously by their respective subject teachers first, as determined by school or district administration, the resulting scores of which are then reviewed anonymously by FMB examiners (lukiolaki [502/2019], §18). Following the exam, subject reports are drawn up in DP, explaining overall rationale for grading as well as marking criteria, while only a general report on the "definitive features of a good response" (lopulliset hyvän vastauksen piirteet) is drafted for each exam in FNC. Schools can subsequently purchase marked exams in DP to help subject teachers understand how students were individually examined (including how internal assessment was moderated), but no such option is available in FNC, and subject teachers are explicitly instructed not to contact FMB examiners to seek their rationale for moderation (FMB 2020d).

The primary written task used to compare the two curricula are paper 1 (DP) and the production task (FNC). Paper 1 for HL English A: Language and Literature, in the 2013 syllabus, is a comparative analysis of two unseen texts (both non-literary or one non-literary and the other literary), with guiding questions that are recommended but not prescribed for student response. Students can opt for two sets of texts to compare and are given 1 hour 30 minutes to handwrite an essay comparing how meaning is constructed in the two texts. The text types used as source material in the most recent four exams include webpages (tourism, fundraising, travel guide), news articles (tabloid and broadsheet), magazine articles, an infographic, a piece of folklore, an oral narrative, a poem, and a blog post. Criteria for the exam use four equally weighted discrete parameters (Table 13).

The primary written task of the FNC matriculation is given in the "production" section but seems to include options for production, interaction, and mediation, based on the definition of those activities in CEFR. The prompts give a specific social context for the piece of text prompted and can be based on unseen or seen texts (if the latter, the text has been used earlier in the exam) that may be literary or non-literary and compulsory or optional to the writing prompt. The questions ask students to process a text, reflect, or argue, requiring minimal content knowledge, in response to the prompt in a word-processed essay, and there may be some recommended music to listen to or images to ponder while composing the written work. The text types used as source material in the most recent four exams include a quotation by a famous figure, an opinion piece, a set of presentation slides, a mind map, and a poem. Criteria for the written text is based on three overlapping criteria, for which communicativity serves as the "primary" criterion (Table 13).

2.4 Hypotheses

The research questions for this master's thesis were given as follow:

1. To what extent do curricula and teaching practices in IBDP and FNC concerning writing reflect the shift from grammar to action-oriented language production, as outlined in CEFR?
2. How valid is assessment of action-oriented written language in both curricula?
3. What are the best practices in both curricula that emphasize the instruction and assessment of action-oriented language?

Hypothesis 1: FNC reflects a larger shift towards action-oriented language production compared to IBDP, (1a) except in more complex forms of production and mediation tasks compared to DP, (1b) in more varied forms of interaction tasks, as a result of backwash, (1c) in its reliance of FonFS instruction.

Based on the research and source background and especially in light of the larger task type coverage in the FNC matriculation exam (Table 5), it is expected that learning, instruction, and

assessment in FNC reflect a substantially larger shift towards action-oriented language production compared to DP. On the other hand, because DP is a CLIL curriculum, it is expected that DP instruction and assessment will reflect more advanced forms of production and mediation. One caveat is that I hypothesize, based on personal experience with students of English in Finland, that mastery CR assessment in the FNC matriculation exam is likely a poor fit for the Finnish classroom, where there is much wider variation of language proficiency than is recognized by the matriculation exam and curriculum, resulting in low criterion-related validity. This would likely mean a poor correspondence of source texts and activities with FNC students, especially those with higher levels of proficiency, resulting in low context validity. Furthermore, my own professional experience with Finnish teachers of EFL/ESL is that there is a preference for FonFS compared to FonF instruction, and the ambiguity.

Hypothesis 2: The FNC curriculum exhibits lower cognitive, context, scoring, and criterion-related validity, (2a) with both curricula scoring low for consequential validity, and (2b) a larger gender performance gap in the IB.

Given the breadth of the FNC curriculum, the exposure of Gen Z adolescents to English outside the school, and what I know of the matriculation exam, it seems to be a hopeless endeavor to me for the exam to capture all the goals of the curriculum as well as account for vast differences in language acquisition, while justifying a mastery CR score based on NR results. My own familiarity with the matriculation exam through FNC instruction and interaction with my teaching colleagues also pre-dispose me to think that the matriculation exam tends to focus on a traditional FonF understanding of language rather than language as being action oriented and context dependent. My experience likewise pre-disposes me to think that there may be a prominent backwash effect in both curricula and that girls tend outperform boys in formal written analysis, the primary form of production assessment in the IB.

Hypothesis 3: Best practices emphasizing the instruction and assessment of action-oriented language draw attention to language as a tool of communication, the noticing of language, and a combined FonF/FonFS approach to language acquisition.

As for the last research question, I do not have a clear hypothesis for what the best practices at the school are. However, based on the research history, it is most likely that they will emphasize the aspects highlighted in the hypothesis.

3 Methods

The thesis uses a mixed methods research (MMR) approach to answer its research questions, using quantitative and qualitative data and methods to inform analysis (see Figure 1, above). More specifically, quantitative data are taken from written production assessment based on the results

from the Spring 2019 exam session (N = 47,880) and from the assessed written work of first-year students for school year 2019–2020 (N = 615). In addition, English teachers at the school studied, excluding me (N = 5), participated in an experiment in which they assessed a student essay used by the IB as guidance for how to apply assessment criteria, the quantitative results of which were used as a small-scale comparison between the two groups of teachers for trends in assessment approaches. The informants also granted access to their Google Classrooms, the primary online learning platform of the school, which contain the material used for instruction and assessment. Observation from these data, together with background research, informed the development of the interview guide (see Appendix B), which was then used to explore teachers' attitudes about and practices in the instruction and assessment of written language.

Given the breadth of the study in the limited space of a master's thesis alongside the practical demands of a professional setting for curricular reform, the study attempts to consider the policy, practice, and research relevant to L2 English learning, instruction, and assessment in pre-DP, DP, and FNC in the institutional context of the school studied. Given the complexity of the research questions and given that much of the data in this research is new or underdeveloped in the field (e.g., pre-DP school curricula in Finland, written assessment comparisons and teaching approaches across DP and FNC), MMR serves the pragmatist purpose of triangulating different data points that are assumed to be complicated by, and therefore need to be treated in light of, various social and individual factors that cannot be extricated from their environment without compromising the applicability of the data; MMR thus helps to weigh out inherent strengths and weaknesses in applying quantitative and qualitative approaches to the data to outline best practices to inform curricular change based on observed problems, evaluating resource- and policy-limited solutions (see chapter 1).

Qualitative analysis serves both an exploratory and a descriptive function that are complemented by quantitative analyses of assessment practices of instructors and institutions (Ivankova and Greer 2015). The interview data offer insight into the practices, attitudes, and perceptions of IB and FNC teachers, which are informed by quantitative analyses of their assessment practices in the experiment and the written assessment of first-year students in 2019–2020, as well as of the IB and FNC exam results for Spring 2019, and by qualitative analyses of differences in the construction of the curriculum and the exams, the latter of which is considered given the backwash effect observed at the school. The primary theoretical framework for the study is action-oriented language instruction (CEFR), a combination of SLA approaches that is used as a touchstone for evaluating best practices in IB and FNC instruction and assessment and used to inform curricular reforms at the school studied. Given that the teaching practices addressed in the

interview are a culmination of a teacher's education and experience in and their continuous reflection on the learning, instruction, and assessment of language, the study thus uses concurrent quantitative and qualitative research analysis to inform the primary qualitative data, the interview with the teachers (see Figure 1).

3.1 Ethical considerations

Following the protocols stipulated in the research permission granted by the City of Espoo on 5 May 2020 and in line with GDPR (EU and Council of the EU 2016, §32), the non-anonymized student data set was drafted in Microsoft Excel under encryption, using 7-zip, then exported to SPSS only after anonymization, using the random number function in Excel. Teacher names were likewise anonymized under encryption, with transcriptions of the interviews made on Microsoft Word in Courier font, then exported to and edited in Notepad++ with ANSI encoding for analysis using RQDA, R package for Qualitative Data Analysis (Chandra and Shang 2017).[3] Transcript conventions follow those outlined in Richards 2003, 173–74, and are listed below, in Appendix A. Because of the ongoing and uncertain nature of the COVID-19 pandemic, as well as to limit unnecessary travel and potential exposure to SARS-COV-2, interviews were conducted and recorded by video call via Google Meet, using City of Espoo–provided G Suites account to ensure data security.

3.2 Sources

This study sought to use a broad range of data as sources for quantitative and qualitative analysis to inform the development of the school curriculum. Quantitative analysis is used to examine assessment practices by teachers and examiners in both curricula as well as whether some of these differences result from differences in demographics, while qualitative analysis is used to examine differences in curricula, assessment practices, approaches to instruction, and learning strategies. Student perspectives were not gathered here, given the volume of data already being examined, as well as logistical and affective issues, especially those arising from mandatory distance teaching, which began on 17 March 2020 for schools in Espoo and continued until the end of the school year for high schools.

---

[3]At the moment of writing, RQDA has been archived from CRAN due to its use of the now-deprecated package gWidgets, replaced by gWidgets2. A workaround, suggested by BroVic (https://github.com/Ronggui/RQDA/issues/38), is to use either the portable version of RQDA (https://github.com/Ronggui/RQDA), or to install RQDA from the CRAN archive (https://cran.r-project.org/src/contrib/Archive/RQDA/) using R (ver. 3.6.3) to access the appropriate dependcies. The .rqda file, including that found in Appendix G, is stored as an SQLite database.

3.2.1 Teachers

Teacher (N = 6)	T00[1]	T03	T01	T04	T05	T02
Curriculum	Pre-DP	Pre-DP	Pre-DP	FNC	FNC	FNC
Essays (N = 477)	46	44	45	156	117	69
Students (N = 159)	23	22	45	52	39	23

[1]*Excluded from the interview, since I am the same person designated here as T00.*

Table 8: Distribution of First-Year Students Across English Teaching Staff

Data were drawn from essay scores by the six English teachers of first-year students for school year 2019–2020, representing the entire English department of the school studied. As I am one of these teachers, care will need to be taken not to overweigh results originating from me among IB teachers and in the English department as a whole. Among the IB teachers, I am the least experienced, with only 4 years of experience in IB instruction, while T01 and T03 have had over ten years of experience across at least two different schools, only one of whom works as an examiner in English A. The FNC English teachers are all tasked by the principal to conduct preliminary matriculation exam assessment and have had between 2 and 5 years of experience in FNC instruction, though only at the school studied, with one of the three having co-authored several EFL/ESL textbooks in Finland. All teachers received their subject teacher training in Finland. The population includes 2 male and 4 female teachers, with three aged 25–35 and three aged 40–50. Given the diversity of the teachers' demographics, the small sample size, the unevenness of course distribution across the teachers and the curricula, imbalanced data exclusion (see sub-sec 3.2.2), known differences across curricula and exam components (see sec 2.1), and likely differences (including pedagogical strategies) in essay assessment, it is not expected that these data will produce statistically significant results with wider implications outside the walls of the school, nor would they be appropriate to use to calculate effect sizes with these data alone. Instead, the figures are used to evaluate validity of assessment criteria and prompt discussion about expected tolerable variation in assessment practices between teachers and across the two curricula.

In addition to the essay data described above, the teachers participated in an experiment in which they were asked to assess the same essay (see Appendix D), having been asked to treat the essay as if it had been submitted by a student in the last course of first-year English. The essay selected was one used by the IB on their online resources to help calibrate assessment, as there are official IB examiner comments and marks published for the essay. The examiner marks were thus

used as a control. An essay for the FNC matriculation exam would have been ideal, but there were no equivalent student essays with examiner marks and comments readily available to FNC English teachers.

Following this experiment, teachers were interviewed for 90–120 minutes via video call, to ensure safe social distancing measures during the COVID-19 pandemic. Due to the close personal relationship between me and the informants, care was taken to allow small talk in the minutes prior and following the interview, to ensure that the interview was treated separately, with some personal distance. On the other hand, the close personal relationship meant that informants were open in their responses, with distance teaching between March and June 2020 ensuring that informants were comfortable using video call as a medium for discussion. The interview guide was developed using incremental complexity, to ensure that informants approached the topic more critically than during the preceding casual discussion, noting differences between curricula, and avoiding pedagogically specific terminology, as an attempt to elicit critical and reflective responses as opposed to responses that draw from pedagogical training rather than from experience (for the full interview guide, see Appendix C).

3.2.2 Students

	FNC	**pre-DP**
Students	114 (*151*)[2]	45 (*50*)[2]
Essays written	342	135
Form groups (FG)	6	2
Students per FG	19.0 (*25.2*)[2]	22.5 (*25*)[2]
L1 Backgrounds[1]	2	13
Swedish as L1[1]	88.6% (*90.7%*)[2]	33.3% (*32%*)[2]
Finnish as L1[1]	11.4% (*9.3%*)[2]	26.7% (*26%*)[2]
English as L1[1]	0.0% (*0.0%*)[2]	15.6% (*18%*)[2]
Other L1[1]	0.0% (*0.0%*)[2]	24.4% (*24%*)[2]
Female[1]	43.0% (*43.2%*)[2]	42.2% (*48%*)[2]
Excluded	26.5%	11.1%

[1]Based on registration in the local registrar, which does not account for number of L1 other than one or gender-nonconformity.
[2]Total figures, which include the students excluded from the overall data set, are given in parentheses. Note that this number includes, for instance, those who did not pass the course, dropped out of school, and transferred between educational lines and who are thus excluded from the tally given in Table 8.

Table 9: Demographics of First-Year Students at School Studied

As the above table indicates, the L1 background of students, which is already more complicated than the table suggests due to bilingualism in families (e.g., for students registered with Finnish as their mother tongue yet attending the Swedish-speaking high school on the FNC line), is much more diverse in pre-DP than among FNC students. Besides including students with English registered as their L1, pre-DP students' L1 languages included Korean, French, Italian, Spanish, Urdu, Bengali, Russian, Somali, Talugu, and Tamil. Some student data were missing from the set and are therefore excluded from the study. The reasons for missing data are various: some were due to students having disenrolled from the online course platform, while others had to do with incompletion of the course or transfers across curricula or schools, the three of which are not necessarily mutually exclusive. Given that the underlying reasons for missing data can include student wellbeing, information about which cannot be ethically included in this study, as well as other personal information without direct relevance to this study, no information is given for why these data were excluded. From a demographic perspective, however, there are few differences among the included variables, except in those related to FNC and pre-DP. As an unintended consequence, since FNC students outnumber pre-DP students, the excluded data produces slightly more comparable results across the two curricula, though FNC students still represent 2.5 times the population size of their pre-DP peers.

3.2.3 Curricula: IB Diploma Program and Finnish High School

Comparison between curricula were made using a mixture of the newer and older syllabi in both curricula. The 2019 DP and 2016 FNC curricula, as well as the digital matriculation exams in English (beginning Fall 2018), were relevant when discussing current teaching and assessment practices, while direct curricular comparisons are made to the 2019 DP and 2021 FNC curricula, given that these are the sources used to develop the school curriculum starting in school year 2021–2022, which is the main goal of this study.

Comparison of FNC and DP exam scores, on the other hand, was made using the results and grade boundaries of the official scores published by IBO and FMB for the Spring 2019 exam session (IBO 2020b; FMB 2020b), using the 2013 DP English syllabus and the digital FNC English exam, since these are the most recent comparable exam results as of the writing of this thesis. Comparison for Spring 2020 was not possible, since these exams were cancelled for the IB due to the COVID-19 pandemic (IBO 2020c). Both exam results represent the older syllabus in the curricula. For the purposes of comparison and in order to align the data with those used in this thesis, only long-form or HL scores were chosen, with only the scores for the Language and Literature curriculum used for DP. As gender data are given only for FNC scores, gender was

omitted from comparison. The most notable caveat to the following data is that the figures represent overall scores in the subject, not just the written component. However, component grade boundaries are given only for DP scores (IBO 2019a), so the score distribution for HL paper 1 is extrapolated on this basis, albeit ignoring how the score of paper 1 is weighed in calculating the overall grade. For FNC scores, the overall subject score is assumed to be equal to the essay score, as no other data are readily available.

As with the preceding data, the lack of statistical rigor in the samples is to be expected and the use of contradictory source material (i.e., old vs. new syllabi) necessary, given the complicated nature of the data and the purpose of this study being action research.

3.2.4 Textbooks and Other Course Content

Given the range of data used thus far, there were limited data gathered to examine the source material used for language instruction. A range of source text types was given, above, in Table 4. As indicated in the table, FNC instruction is largely reliant on textbooks for source texts and general instruction, while DP instruction does not use textbooks.

In addition, the learning platform (viz., Google Classroom) and shared teaching material (via Google Drive) were used to offer cursory analyses of and comparisons between teaching practices and learner strategies in the first-year English courses, with permission granted by the teachers. It should be noted that the third course, and some of the second course, of English was heavily impacted by changes necessitated by mandatory distance teaching.

3.3 Coding Themes

Coding themes for qualitative analysis derived primarily from CEFR and the interview data, which was done using RQDA, based on the research background, including pedagogical strategies. These themes were, in turn, applied to the instructional, curricular, and assessment materials (viz., course essays and other writing tasks, as well as textbooks and other course content). In all, eight themes were identified, a breakdown of which is given in Appendix D and the application of which to the interview data can be retrieved from the SQL file located in the research data repository (see Appendix G): assessment, validity, communicative language activities and strategies, communicative language competence, plurilingual and pluricultural competence, approaches to teaching, approaches to learning, affect, and challenges and opportunities.

While almost all codes used in each theme could be drawn from CEFR or the research literature, four codes proved difficult to fit into the literature review, all from the same theme of

communicative language competence: communicativity, register, style, and politeness. The latter three were divided here simply because they presented interesting differences in the instructional context studied, whereas they are presented under the single heading of "sociolinguistic appropriateness" in CEFR.

Communicativity, on other hand, proved to be an ambiguous term, at times seeming to be used to refer only to linguistic competence, at times to general comprehensibility, and at times to what are described in CEFR as communicative language competence as a whole. As such, communicativity was coded as a sub-theme of its own, since the use of the term is specifically related to the FNC matriculation criteria, as a direct translation of "viestinnällisyys/Kommunikativ förmåga," and it could not always be ascertained from context how the source was using the term. Thus, the code was only applied to the interview transcripts and FNC curricular and exam material where the term was used explicitly.

4 Results

For the purposes of this section, action-oriented written language, based on the discussion in sub-section 2.1.2, is defined as language use that may be

1. relative to a social, cultural, and/or generic context,
2. relevant to an attempt to negotiate for meaning or as a tool for communication, and
3. can be assessed based on
    a. communicative language strategies and activities of production, mediation, and interaction;
    b. linguistic, pragmatic, and sociolinguistic competences (as opposed to only linguistic competence);
    c. plurilingual and pluricultural competence.

The following sections discuss the evidence of action-oriented language in the curriculum and course and exam assessment in both educational lines (4.1), comparing these findings to how action-oriented language were actually assessed in the first-year English courses and in the sample essay assessed by each informant (4.2).

4.1 Curricular Evidence

To offer a fair comparison of the two curricula, since the updated CEFR reference levels were only published in 2018, curricular analysis focuses on the 2019 syllabi for the two curricula, whereas analysis of curricula in practice necessarily focuses on the 2016 syllabus for FNC and the 2019 syllabus for IB, given that these were the syllabi used for first-year English instruction for school year 2019–2020. The same syllabi are used to discuss the Spring 2019 exams, the most recent comparable evidence for both curricula (cf. Table 3). The following section thus surveys the

evidence in the syllabi (4.1.1), then turn to how the curricula were realized in instruction (4.1.2) and in the May 2019 exams (4.1.3).

4.1.1 Curricula

The FNC and the IB curricula present different approaches to the instruction, learning, and assessment of action-oriented, based on differences in the expected language profiles and end results. As shown in Table 1, the target proficiency of English A in FNC is B2.1, whereas the target proficiency of English A: Language and Literature is B1.2–C2. Given that the CEFR benchmarks are given for language as being action oriented, it is expected that the curricula should reflect action-oriented language instruction, such that the definition of action-oriented language should be reflected in the curricula.

In the first case, to do with the social, cultural, and/or generic context of the communication, both curricula do approach language as communication in context. In the case of FNC, the English language is explored through six mandatory (1–6) and two optional (7–8) modules:

	Name	**Social**	**Cultural**	**Generic**
1	Study Skills and Language-Identity Construction	Language profiles, study skills (self-understanding), transactional language		
2	English as a Global Language	Language in nations and the international system	English as a Lingua Franca (ELF)	Communication styles in different forms of media
3	The English Language and Culture as a Tool for Creative Expression	Language in the creative sphere	Cultural topics important to the student	Language of and about creative expression
4	The English Language as a Means of Influencing	Language in civil society		Persuasion in media and self-expression
5	A Sustainable Future and Science	Conducting research and other forms of inquiry	English as a scientific language	Source reliability (popular vs. scientific texts), simplifying and explaining texts
6	The English Language in Higher Education and Working Life	Language as social capital (for studies, work, and related situations), in working life and international contexts, in national or international organizations		Formal language
7	The Environment and Sustainable Living	Global citizenship, environmental issues and solutions, international organizations and negotiations		Forms of discussions about environmental issues, develop skills for source criticism
8	Communication and Persuasion in Speech	Effects of different English language profiles	International contexts for spoken English, variants of English	Factors of oral interaction

Table 10: Social, Cultural, and Generic Context in the 2021 FNC English Curriculum (Opetushallitus 2019, 180–85)

In the IB, on the other hand, teaching is divided evenly among three different course themes, which can be divided up and arranged by the teacher, with certain requirements separate from individual course themes:

	Name	Social	Cultural	Generic
1	Readers, Writers, and Texts		Contexts and complexities of production and reception	The nature of literature and its study (i.e., textual or rhetorical features; stylistic, rhetorical, and literary elements); nature of language and its communication (i.e., authorial choices made by authors of words, image, and sound)
2	Time and Space	The impact of literary and non-literary sources on and by the social or political environment, historical perspectives of non-literary and literary sources, cosmopolitan nature of texts	Context of language use, (broadening of) personal and cultural perspectives on the context of meaning, cultural perspectives on non-literary and literary sources	
3	Intertextuality	Language as a system of communication		Thematic concerns, generic conventions, modes or literary traditions
Other Requirements		Audience and authorship of literary and non-literary texts	2–4 works (of 6) in translation; works from at least three separate centuries; authors from at least three different countries, across two different continents	Works from three of four literary forms (fiction, non-fiction, poetry, drama)

Table 11: Social, Cultural, and Generic Context in the 2019 IB Language A Curriculum, HL (IBO 2019b)

Based on this comparison, the FNC English syllabus clearly has a larger and broader focus on various social contexts of language, relating directly to everyday life and agency in politics and civil society, whereas the IB curriculum focuses especially on the study itself, rather than the practice, of language and literature. On the other hand, the IB Language A syllabus has a broader approach to cultural context of language use, including the access of other cultures through translation, whereas the FNC curriculum focuses on exploring the cultures of differing manifestations of and context for English. Both curricula specify that these goals should be informed by regular formative and summative assessment, including written tasks that develop the skills towards the above goals (IBO 2019b, 31; Opetushallitus 2019, 45).

The two curricula employ different strategies to foster an approach to written language that emphasizes the negotiation for meaning, though the FNC English A syllabus seems to do so more explicitly and broadly. Group work is emphasized throughout the syllabus, for instance, as well as differing forms of constructive interaction across different media, particularly in the second module.

Opportunities for meaning-negotiation in writing are prevalent especially in modules 2, 4, and 7, which focus on meaning-negotiation on various media platforms, as well as in international legal contexts. The curriculum also emphasizes how language should be used "väline- ja taitoaineina" (Opetushallitus 2019, 174), as tools for self-understanding, skill development, critical and ethical thinking, and other interaction with the world. The goals of the syllabus focus on the ability for students to become autonomous in their language learning and be able to use language effectively in a variety of contexts, rather than to demonstrate mastery of language as such, a model that coheres well with CEFR.

In exploring language as a tool for communication, the IB syllabus is much more limited. Written forms of meaning-negotiation can manifest in practice exam papers and the HL analysis essay, but these focus on meaning-negotiation in an academic setting—what is meant in argumentation and interpretation specifically rather than a broad context of communication. Some possibilities for reflection on meaning-negotiation in writing are offered in the form of the Learner Portfolio (LP), a collection of the student's work stretching across DP that is not assessed but which may be used to demonstrate student progress in cases where academic dishonesty or maladministration are being investigated. Here, the syllabus explains that the LP may be used to reflect on, among others, negotiations that take place in "classroom or group discussions" and experiment "with form, media and technology" (IBO 2019b, 26), but these seems to be largely productive, receptive, and mediational in nature rather than interactive. Like the FNC syllabus, the IB syllabus does not focus on mastery of language as such but rather the mastery of skills associated with the study of language and literature. Unlike the FNC syllabus, the goals are largely receptive, such that the student would be exposed to a broad variety of texts and be able to engage them in mediation and production but not necessarily in interaction. Another interesting departure from the FNC goals are the IB's affective goals, that students be able to communicate "confidently" and develop a "lifelong interest in and enjoyment of language and literature" (IBO 2019a, 14).

As a tool for communication, language is thus studied much more widely in the FNC than in the IB syllabus, with broader contexts for interaction and production being the key advantage of the FNC syllabus. The emphasis in the IB on developing confidence in and the appreciation of language and literature is interesting, however, and brings an added focus to motivational and affective matters, which are also relevant to language learning.

While both curricula offer a broad understanding of how English should be taught as an action-oriented language, neither curriculum specifies how such skills, and writing skills in particular, should be assessed. These are realized instead in the classroom and on exam days.

4.1.2 Curricula in Practice

Writing assessment in both curricula largely followed the differing emphases on production, interaction, mediation, and reception for school year 2019–2020. In FNC classes, students were exposed to a wide variety of writing contexts for a wide variety of largely non-literary texts, while, in pre-DP classes, students were exposed to a limited variety of writing contexts for a wide variety of literary and non-literary texts. Furthermore, FNC instruction followed an explicit FonF/FonFS model of language instruction, which shifts towards explicit FonFS models of assessment, whereas pre-DP instruction used an implicit FonF model of language instruction, with an assessment model that focused on implicit forms of language mastery.

In FNC instruction, explicit FonF/FonFS language instruction manifests in its structural, proactive approach to language, where linguistic competences are largely developed intentionally rather than incidentally. This approach is largely traditional, in that the first three courses offer a broad overview of grammatical aspects of language, supported incidentally by authentic texts reproduced in textbooks:

Extract 1 *T02:1048–53, Structural FonFS Language Instruction in FNC*

T02 like, we have this- uh, (x) ↑ pseudo-curriculum, which is, uh, our material- and, uh, yeah- of course, that, (..) mm, that affects a lot what- like, (x) what it says in the textbook (.) is sometimes what we- (..) FOR EXAMPLE, IN TERMS OF grammar, that's what (x) we just have (.) with them, even if it doesn't say in the curriculum

An FonF approach to instruction is realized in the textbooks by focusing on texts first to investigate areas of communication breakdown before turning to forms to foster conscious rule-formation. A structural approach is also taken in the instruction of written production skills:

Extract 2 *T05:954–963, Structural Instruction of Written Production in FNC (2)*

T05 course one, it would be just general, kind of, what is an essay- matriculation exam, what is an essay- ((laughs)) matriculation exam. Course two, we are looking specifically at cohesive ↑ markers, so how to guide the reader- um, linking words, all of that- and, also, paragraph breaks- if they don't have it down by then, uh, in course three, um, <I've opted for letting them have a more, uh,> creative task, but also kind of playing, if they want to, with, kind of, uh, style, with register- there, I, um, kind of, encourage them, if they want to, to, uh, take on the role of how an eighteenth-century, nineteenth-century (.) person (x) <would write a letter,> and so on.

As noted in Table 4, FNC teachers also employ a broad range of productive and interactive writing tasks, most of which are self- and peer-assessed, given the volume of work that would be needed to provide teacher assessment (T02:627–36).

Assessment components in FNC are almost entirely based on forms of recall that rely on explicit FonFS models of language instruction. Namely, course quizzes entirely are tests largely are made up of word completion, word/phrase translation, open-ended cloze, sentence correction, and

multiple-choice responses. Some tests also included short sections of continuous writing, usually as bonus points, fitting the theme of the course. Such tests are clearly geared towards efficiency of objective assessment, targeting conscious rule-formation and the automization of language acquisition.

Students are also required to write a short essay in each FNC course, the prompts and criteria deriving from the digital matriculation exam. The use of essays for written language assessment seems to be in line with an FonF model of language acquisition, prompting students to produce a text that would seem to prioritize language as message instead of as code:

Extract 3 *T02:114–37, Essays as FonF assessment in FNC*

T02 I mean, at least< (x) for ↑ me, like, (x) for English- for first-year
students, (x) I try to, sort of, grade the essays (.) mmm ((dental click)) so they reflect (.) what we've been ↑ doing (.) during the course .hhh for example, now, °in English 3, I've° (.) paid more attention to ↑ punctuation, because we had, °like, punctua°tion. Um, and so I would probably- if I had (.) an English (.) A (.) course (x) for pre-DP, I'd probably, like, try and find
what they have studied, °because it's a course where basically-°

IR Yeah, and I appreciated in the comments, you kind of, like, (.) alluded to that, you know-

T02 Yeah, I would have, yes ((laughs)) .hhh and I assume that you've been (.) learning that, and, I mean, (..) um, hhh ((dental click)) .hhh for me,
until like- >actually, all that course of English that I< <now teach,> um, the essay's not- (.) I mean, the (x) main purpose of the essay is not to (.) mm, prep the students for the matriculation ↑ exam, but (.) to actually apply what we've learned, you know, so it's not like (.) I don't want to- also, for them, I don't want them to think, like, >okay, now< we're <from first course->
because, you know, we start writing the essay .hhh you know, from the first course, I don't (.) <want them to think of the actual> matriculation exam, but, like, a way to (.) ↑ apply what we've ↓ learned.

IR Mmhmm, is that (x) common (.) ↑ between all the English teachers? Or, is that-

T02 MOSTLY, yes. <I think this year I've been doing that> more .hhh because of the pilot ↑ project .hh so we've been, like, hh I've been (.) maybe more ↑ conscious (.) of ↑ that

At least one focal point of matriculation exam essays, as used for achievement assessment, is thus to assess students' ability to produce meaningful communication using the forms and contexts explored in a course. This approach also includes an incremental process of developing task familiarity (Extract 2). Both aspects of essay assessment thus take structural approaches to language instruction but follow an action-oriented approach is applying them to a specific context, using a constrained-constructed response format, combining the advantages of FonF and FonFS approaches to language acquisition (in addition to the primary disadvantage: subjective assessment).

In pre-DP instruction, on the other hand, implicit language instruction manifests in its task-based, reactive approach to language, where linguistic competences are developed incidentally rather than intentionally. The following table summarizes how the written activities as realized in both curricula compares with both the curricula and the 2018–2020 exams:

		FNC	*DP*
Production	Creative Writing	✓	
	Written reports	—	
	Essays	✓	✓
Interaction	Correspondence	✓	
	Notes, messages & forms	—	
	Online conversation & discussion	✓	
	Goal-oriented online transaction & collaboration	—	
Mediation	Relaying specific information in writing	✓	✓
	Explaining data in writing	✓	✓
	Processing text in writing	✓	✓
	Translating a written text in writing	✓	
	Note taking	—	
	Expressing a personal response to creative texts	✓	—
	Analysis and criticism of creative texts	—	✓

Green: present in course assessment but absent from 2018–2020 exams
Red: present in 2018–2020 exams but absent from course assessment
Purple: present in curriculum but absent from both 2018–2020 exams and course assessment

Table 12: Written Task Types in DP and FNC Exams, 2018–2020, Compared with Course Assessment

One notable departure from FNC teachers' largely procedural approach to written assessment is in their approach to formative assessment of essay-writing when compared to their IB counterparts. Namely, FNC teachers color-code mistakes by nature of the error (i.e., orthography, punctuation, syntax, verb-specific errors, preposition/pronoun-specific errors, redundancy, register, and contextual appropriacy). The color-coding guide is also written using non-technical language—for instance, inappropriate register is color-coded as "chatty language." In the sample essay, FNC teachers also posed questions where communication was unclear in the writing, using non-technical language. IB teachers, on the other hand, corrected mistakes in the sample essay directly, commenting largely on the quality, cohesion, and thematic development of the analysis and only minimally on the language. The only comments on sentence-level errors by one of the IB teacher also employed technical language in noting sentence fragments in the sample essay.

This comparison indicates that, at least in formative assessment, FNC teachers rely on and foster student's ability to notice gaps in communication, in line with Schmidt's (1990) understanding of language as a tool for noticing communication breakdown and construction, whereas IB teachers rely on students to access explicit rules of language. In both cases, formative

assessment is disconnected from instruction. That is, pre-DP instruction follows implicit FonF language instruction but, in formative assessment, relies on FonFS language recall, whereas FNC instruction follows explicit FonFS language instruction but, in formative assessment, relies on FonF language negotiation skills.

4.1.3 Exams

	DP, Paper 1 (English A: Language and Literature HL)	*FNC, Written Production (long-form English)*
Number of criteria with action-oriented descriptors[1]	2	2
Weight	50%	N/A
Description of criteria (action-oriented aspects)	**Language**: clarity, effectiveness, carefulness in choice and precision, accuracy (in grammar, vocabular, and sentence construction), appropriacy and effectiveness in style and register **Analysis and Evaluation**: understanding of textual features and/or authorial choices, evaluation of how those shape meaning **Understanding and Interpretation**: understanding of the literal meaning of the text, convincing interpretation of its implications and subtleties, well-chosen and effective references to the text	**Communicativity**: clarity, authenticity, smoothness, nuance **Breadth and Accuracy of Language**: breadth, variety, idiomaticity, contextually appropriacy, general excellence
Other criteria	**Focus and Organization**: organization, coherence, focus	**Content and Structure**: variety in treatment of topic, personal style, logical flow, use of cohesive markers

[1]*Action-oriented language here simply denotes aspects of written production and mediation that focus on meaning instead of communicative linguistic competence.*

Table 13: Criteria in FNC and DP Focusing on Assessment of Action-Oriented Language (based on IBO 2019b, 45–53; FMB 2017, 16)

	FNC	DP
Total number of long-form / HL English exams taken	19,995	27,885
As a percentage of total English exams taken	97.79%	61.87%[1]
Written composition as a percentage of overall score	36.45%[2]	70%
Median score	4	5
Mean score	4.14	4.95
Min. passing score	2	3
SD	1.486	1.006
Min. passing score (σ)	m − 1.44σ	m – 1.94σ
Failure rate	4.26%	0.60%
Skewness	-0.096	-0.148
Kurtosis	-0.601	-0.147

[1]This figure reflects the proportion of HL students only in English A: Language and Literature.
[2]This figure includes both the translation exercise (task 15) and the essay (task 16).

Table 14: Exam Score Comparison of Spring 2019 Exams in FNC and DP Long-Form/HL English (drawn from IBO [2011] 2013; 2020b; FMB 2019; 2020a; 2020b)

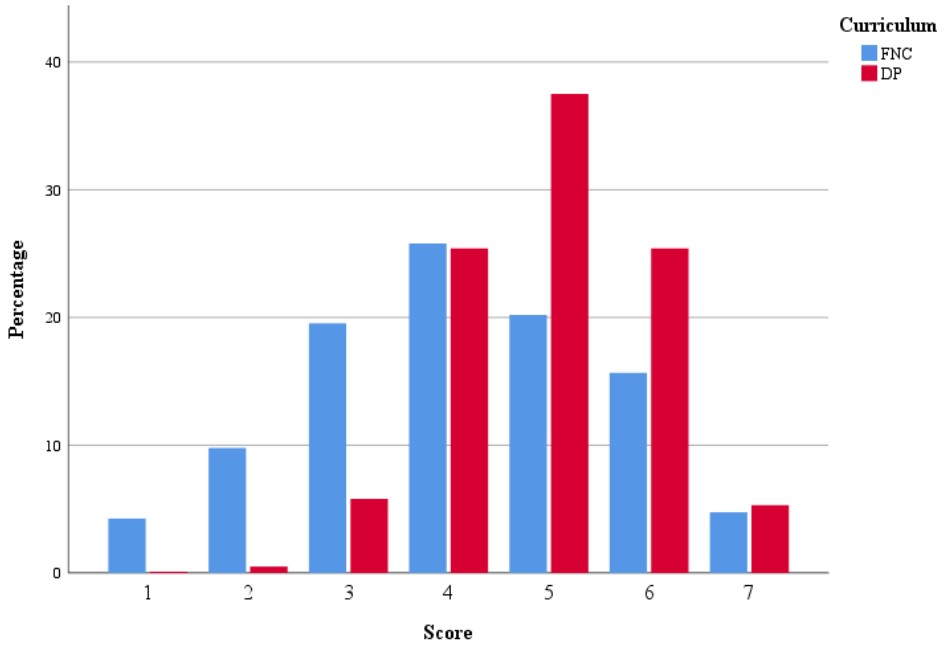

Figure 3: Exam Score Comparison of Spring 2019 Exams in FNC and DP Long-Form/HL English (drawn from IBO [2011] 2013; 2020b; FMB 2019; 2020a; 2020b)

The data show that exam scoring for both curricula skews exam results above than the mean compared to below, with a minimum passing score given at approximately ½ σ lower than the mean in DP compared to FNC. This is a significant difference, as reflected in the higher failure rate of students examined in FNC, but unsurprising given that FNC exam scores are graded on a set curve, whereas IB exam scores are set according to correspondence of component grade scores with descriptive criteria (see sec 2.1). Unfortunately, no data are available for the exam grades of IB English students in Finland. Assuming, however, that the distribution of grades do not differ

between Finnish and non-Finnish students of DP English, the results indicate either that FNC and DP EFL education result in substantially different learning outcomes or, more likely, that there are flaws in how the exams are assessed and, subsequently, how those exam results are used for CEFR benchmarking.

4.2 Validity

Due to space limitations, I am only to focus my analysis on scoring validity here (i.e., the validity of assessment based on how assessment is applied by scorers). This form of validity represents the most immediate form of validity that can be addressed in curricular change at the school level, while most other forms of validity would require structural changes in the curriculum that go beyond the scope of changes possible in a single high school (4.2.1). I then proceed to offer a cursory review of other forms of validity relevant to the data gathered (4.2.2).

4.2.1 Scoring Validity

For the scoring criteria to exhibit scoring validity at the rating scale level, the criteria used must be discrete in nature. In other words, the criteria should (1) measure separate abilities and (2) distinguish between levels of performance, and (3) spread test-takers into the different levels intended by the test (Knoch and Chapelle 2018, 484–87). To demonstrate that the criteria measure separate abilities, criteria should exhibit slight (0.20–0.35) or limited (0.35–0.65) bivariate correlation with one another, using Pearson's *r* (Creswell 2013, 347). To demonstrate that a test distinguishes between different levels of performance, a wide range of the criteria should be used, though this is dependent on the skill level of the sample population. Finally, to demonstrate that the test-takers are separate into levels intended by the test, the overall score should be determined as intended by the institution.

To test whether separate abilities were measured in the pre-DP essay (N = 113) scores, the scores for the four criteria used to mark each essay were first transformed to be of equal range, then a Pearson's *r* test was run for the four criteria, giving the following results:

Criterion	**A (Understanding and Interpretation)**	**B (Analysis and Evaluation)**	**C (Structure)**	**D (Language)**
A	-			
B	0.567**	-		
C	0.432**	0.251**	-	
D	0.480**	0.509**	0.442**	-

*** Correlation is significant at the .01 level (2-tailed).*

Table 15: Correlation Matrix for pre-DP Essay Criteria Scores (2019–2020)

As the above table demonstrates, the criteria meet the 0.65 threshold in all six cases, with criteria B and C demonstrating only slight correlation. However, when the same test was run separately for each of the three pre-DP teachers, the slight correlations (T01: corr([A D] [B D] [C–D]) did not achieve statistical significance, and, of the remaining 15 correlations—all significant at the .01 level [2-tailed] except at T01: corr(B C)—only T01: corr (A B) exhibited moderate correlation (0.66–0.85), with an *r* value of 0.659, at the very lowest edge of the parameter. All other cases exhibited slight correlation. It should be noted, however, that, due to incomplete data recording, criteria grades were available for only 23 of the 46 essay grades for T01, whereas those for all 44 essay grades for T02 and all 46 essay grades for T03 were available. Given that all the variations in the data between comparisons of the data across and by the teacher evaluator all derive from T01, the variations are likely the result of sample size differences. The slight and limited correlation of the pre-DP essay criteria scores found here are echoed in the interview data:

Extract 4 *T01:118–28, Rating Scale Validity of IB Criteria, Separate Abilities (1)*

T01 You know, if there's something- if there's a weakness, like, with the language, it doesn't automatically mean that the ideas are weak, (.) so
120 learning how to um (.) to separate those things and identify- and of course (x) >of course the way that the IB criteria are set up, too, is that they're not kind of< like a ↑ punishment, they're like a- (.) you're (.) awarded for what is there and not what isn't. So um (2.0) [but I]
IR [so did] you think that this
125 criterion separation- is that a strength of the IB's [criteria?]
T01 [I think] so, yeah. I think so, yeah. Once (x) once you learn how to (.) use them. It (x) it's hard to learn how to use them.

Extract 5 *T03:216–23, Rating Scale Validity of IB Criteria, Separate Abilities (2)*

T03 When I was giving the IB grade, I was- (x) the language is (x) a pretty minor consideration, when I- when I'm reading and assessing IB work. (x) As I read a paper one, (x) my brain, as I go through it, is primarily thinking about A and
220 B (. . .) criteria, and then C and D are largely an afterthought. (x) I don't really think about C and D until I've finished reading. Whereas when I'm reading, I'm constantly, (. . .) uh, constantly evaluating the A and B criteria.

T01 explains that each criterion should be determined independently of one another, with T03 emphasizing that criteria A and B are at least determined independently from criteria C and D. Curricular information echoes T01's claim, in recommending that "Good practice with [criteria marking] is to make sure that the criteria are independent of each other" (IBO 2018, 87). Given the quantitative and qualitative data, it is thus highly likely that the IB criteria are used at the school studied as discrete criteria to measure separate abilities.

Correlational analysis on the quantitative data for FNC scores could not be conducted, because the matriculation exam score is holistic, even though analytical criteria are given. This

contradiction was reflected in the interview data, where informants suggested that their assessment of each criterion tended to impact assessment of other criteria:

Extract 6 *T02:799–808, Rating Scale Validity of FNC Criteria, Separate Abilities (1)*

T02 this essay also had (x) some instances, and, uh, for ↑ me, that's (.) something that, uh, really affects the readability, (..) because then you have to- (x) the reader (x) shouldn't have to go back, like, (.) HMM, wait a minute, was this actually a singular? (x) °or a plural° (..) thing, like, (.) so. That's- becau- (x) in this (.) communicativity (.) ↑ box, (.) um, the most tangible thing is, sort, that I ((dental click)) usually, sort of, ↑ first look ↑ into (.) is how easy it is to ↓ read (.) the text. And if you have to go back- (x) it's that- what I say- it's the (x) rule I will (x) say to my students, .hhh if I have to go back (.) a lot, (x) in your sentences, like (.) okay, where was the subject, <then that drops (.) those points (.) a lot,>

Extract 7 *T04:236–44, Rating Scale Validity of FNC Criteria, Separate Abilities (2)*

T04 especially with students that are (.) really ambitious, (x) and <are at a high level,> and (x) very kind of, um- and who might, already in the first year, be thinking about the matriculation exam- hh to those, I tend to stress the point that (.) (x) communication, communication, because, uh, <those kinds of students tend to, um, commit errors that, uh,> that take away the understanding, because they, um, make it too complicated, (x) or, um, experiment with really complex ↑ structures, which I want them to ↑ do, but then (x) if their meaning is lost, then, it- yeah- (x) they fall (x) by quite a few points.

Extract 8 *T05:56–62, Rating Scale Validity of FNC Criteria, Separate Abilities (3)*

IR And do you see (x) that communicativity part as being all three of what you just said. Like, the (.) spelling, the grammar, and the- (..) how well it's written, or (x) where does that aspect come in-

T05 Spelling, grammar, but, then, <underneath that,> the whole structure- so, firstly, the (basic) structure, but then, also, the logical flow (.) through the argument, so (x) the use of linking words, the use of an obvious (.) red thread going through it- all those things.

From the interview data, it seems clear that all informants treated the three criteria as interrelated, given that communicativity could be impacted by morphological (T02:803–804) or syntactic errors (T04:240–241), which have to do with the criterion "Breadth and Accuracy of Language," or by problems with coherence and cohesion of the essay (T05:59–62), which has to do with the criterion "Content and Structure." The interdependence of these criteria does not, however, present a major challenge for holistic assessment, given that holistic assessments are not usually produced with a set of criteria that purport to measure separate abilities. The problem here, however, is that the criteria are presented in the matriculation board material analytically, as three discrete skill sets, with points on a scale of 0 to 99 used to describe eight levels of achievement. The instruction above the criteria, that "overall evaluation of written performance is made primarily based on the criterion 'Communicativity'" (FMB 2017, 16), further suggests that the rating scale is meant to be used as analytical criteria.

To test whether the FNC and IB essay criteria are used at the school to distinguish between levels of performance, the data for both curriculum were collated to facilitate comparison. In the case of the pre-DP essays, the four criteria were set out of 5 points, corresponding to the 5

achievement bands given for the paper 1 HL criteria. In the case of the FNC essays, achievement bands were calculated from the data based on the achievement bands given for the long-form matriculation exam criteria, treated as holistic assessment grades, since component grades were not given.

AB	**Pre-DP (N = 452)**		**FNC (N = 342)**	
	Count	%	Count	%
1	7	1.5%	0	0%
2	16	3.5%	0	0%
3	74	16.4%	0	0%
4	207	45.8%	4	3.5%
5	148	32.7%	37	10.8%
6	N/A		213	62.3%
7	N/A		88	25.7%

Table 16: Frequency Table for Achievement Bands (AB) of First-Year Students' Essay Grade, 2019–2020

Without external measures to serve as controls, it is impossible to say whether the distribution given in Table 16 is accurate to the skill level of students and thus to judge whether or not the scale is used to distinguish between real levels of achievement, especially considering that these essays were used as assessment components within a course rather than as continuous achievement assessment as opposed to fixed-point proficiency assessment the criteria were designed for. Thus, it is not improbable that lower achievement bands were used less frequently and higher achievement bands more frequently to promote a growth mindset, a strategy that several teachers alluded to:

Extract 9 *T04:782–86, Application of Achievement Bands to Promote a Growth Mindset (1)*

T04 So I'm supposing this is a journey that we're going through, over the three years that I'm teaching them. So, maybe, in the first year, um, I would say that the score I gave this was on the high end of what I would have scored
785 it, cause it was a first-year student, and I've said before, that I tend to mark my first-year students a bit higher. By the third year- (x) if I had received that essay from a student in the third year, they would have got below 70%, definitely.

IR Okay.

T04 They would have got, probably, 68, (.) because I couldn't have allowed them to have written that badly.

IR ((dental click)) So you (x) use that kind of grading level to, kind of, (x) encourage- motivate students.

T04 Yes, very much so, the first year, yes, I would have done.

Extract 10 *T01:1069–79, Application of Achievement Bands to Promote a Growth Mindset (2)*

T01 with pre-DP, (.) when we've got .hhh
1070 kind of (.) a hybrid, I guess, assessment (.) going on (.) in our (.) ((dental click)) pre-DP English courses that we've taken some, you know, concepts or, um, <assessment criteria (..) into the (..) into the pre-DP English course

assessment as well.> But it's not as (.) um, (.) or, at least for me, it's not as demanding. (3.0) um (4.5) but then as soon as you hit the- (. . .) and then
1075 hhh I don't know if I really consciously do it, but, like, you know, (.) in the beginning of IB1, even you're like, okay, well this is your first paper one, I'm not gonna, like, ((dental click)) (. . .) I try to be- to use the- I try to use the assessment (x) >crtieria consistently, I'm not saying that you- < I get ↑ harder at using it, but then my- maybe my expectations (.) grow.

For both FNC and pre-DP, teachers thus linked their philosophy of how students might grow over time to how they apply the criteria. The lack of the lowest bands used in the FNC is still surprising, however, given that the population is first-year students, so the concentration of 88% of FNC students at the highest 28.6% of the achievement bands is striking compared to the 32.7% pre-DP students at the highest 20% of the achievements, or 78.5% at the highest 40% of the achievement bands. Interestingly, the other IB teacher interviewed and I take a different approach to promote a growth mindset, as reflecting in the following extract:

Extract 11 T03:1056–63, Application of Achievement Bands to Promote a Growth Mindset (3)

T03 in the old days, when I was doing IB1, I used to (.) consciously give them, (.) like- yeah, one mark extra (.) in each criteria, >I can't remember what it was,< because, otherwise, they'd be getting like- you know, the weaker student would be
1060 getting, like, 8 out of 20, or something, (..) which feels a bit (.) demotivating. But this actually turned out to be (.) even more demotivating, (2.0) because (.) they would get better, (. . .) but their marks wouldn't get better.

This approach of promoting a growth mindset by applying the criteria without correcting for the student's experience in the test format or education in the subject explain the data in Table 16, given that 100% of the scores given in the lowest two achievement bands derive from essays marked by T00 and T03.

Teachers thus accessed the entire range of the criteria in the case of pre-DP, whereas, in the case of FNC, teachers only used the highest four achievement bands, with 96.5% of the grades deriving from the first three of the seven. One important caveat to this comparison is that, from the pre-DP data, all of the marks given at the lowest achievement band and 14 of the 16 marks given at the second-lowest achievement band come from essays I myself graded (T00). These marks were given before the thesis was designed, however, such that their anomalous appearance in the data should be seen as a characteristic of mine as a rater rather than a conflict of interest in the data.

As for the capacity for the essays to spread test-takers into different levels as intended by the test, this aspect of scoring validity can mainly be assessed by examining how the criteria are used as NR or CR assessment and the extent to which these correspond to CEFR descriptors, explored above in sub-section 2.3.2. Additionally, it is also worth investigating the degree to which essays are marked as intended. In the case of the IB essay, the criteria are analytic and evenly weighted, whereas, in the case of the FNC essay, the criteria are analytical with the first criterion given the largest, albeit unspecified, weight, yet the score given is holistic.

If the FNC essays are marked as intended, then there should be an obviously larger effect size exerted by the criterion "Communicativity" over the other two criteria. Unfortunately, because only the holistic score is given, teachers and examiners do not usually assign individual criterion scores to essays marked. In the sample essay used to examine rater characteristics, teachers were thus asked to record individual criterion scores in addition to the overall score marked.

	T02	**T04**	**T05**
Communicativity	78	68	78
Meaning and Structure	88	68	78
Breadth and Accuracy of Language	90	80	80
Mean, equal weight	85.33	72	78.67
Mean, 50%/25%/25%	83.5	71	78.5
Mean, 60%/20%/20%	82.4	70.4	78.4
Overall Score, given	85	72–75	78
Overall Score, next-lowest	82	70	75

Table 17: Criteria Grades for Sample Essay, FNC Scores

While it is not possible in a sample size of 3 to extract a statistically significant correlation matrix for the criteria, it is interesting to note that the overall essay score for this sample seems to have been determined by the criterion "Breadth and Accuracy of Language," given that this criterion was rated the highest by all three teachers and the overall score was given at or slightly above the mean for two out of three of the teachers. In the data given in Table 17, with the most conservative estimate, the criterion "Communicativity" could not have been weighted more than twice as much as the other two criteria in the case of T02 and T04, whereas, in the case of T05, "Communicativity" could have been the only criterion used to determine the overall grade. Given that the criteria state only that "overall evaluation of written performance is made *primarily* based on the criterion 'Communicativity'" (FMB 2017, 16, emphasis added), as well as the small sample size here and the use of an IB essay by an unknown student instead of an FNC essay by a known student, these data can only be taken as a suggestion that the FNC essay criteria may be inappropriately focused on the criterion "Breadth and Accuracy of Language." This suggestion is also corroborated by interview data:

Extract 12 T03:76–100, Communicativity as Other than the Primary Criterion for FNC Essay Assessment (1)

T03 (x) And the national side, (..) despite what the marking criteria you gave (.) us to look at (..) claim, I- (x) my impression is that- (x) certainly, back when I was doing it fifteen years ago, uh it would be- essays were not marked >according to those criteria. I don't think if they were supposed to be, but they certainly weren't. And I don't think much has changed now. I don't think English teacher on the ground (..) use those criteria (x) that you produced. Where did they come from? Are they the real criteria, or?

IR Those are the actual criteria provided by the (.) Finnish Matriculation Board.

T03 Yeah, (. . .) my- I mean, (..) I did mark some essays- >like, last year, I did a couple of English 7 courses, and I marked some essays, and I,

kind of,< (x) I asked (xxxxx) or somebody else, I can't remember who it was- (xxxxx) or (xxxxx)- (x) to look through a couple of essays with me (x) and see if I was (x) still on, more or less, the right lines, and we (.) (x) we seemed to be, kind of, more or less, the same, so- (x) I got the impression that >things hadn't changed that much. And it's much more about just looking at how they write.< (3.5) Uh, I mean I hhh thing- things that stay in my mind from (xxxxx) were my (.) experienced colleagues telling me stuff like, (.) if the verb doesn't agree with the subject, then they're not gonna get more that seventy points. And this essay included (.) some of those (..) for example, the audience (. . .) um, (x) with a plural verb, or whatever. (..) Or I (x) can't remember exactly what it was- but something like that. Um, (2.0) so, hhh in my experience, national-side marking tends to be more (..) prescriptive °in its, sort of,° language, and .hh (x) much more limited to saying- to seeing, have you learned to do this in English yet

Extract 13 T02:661–66, Communicativity as Other than the Primary Criterion for FNC Essay Assessment (2)

T02 many of them first, uh- during the first courses, they (x) try to (.) write something, like, (.) very ↑ complicated, and then (.) then that's- (..) you know, >that destroys the communicativity, because (x) they might< .hhh use, um- I don't know, (x) shortened clauses, (x) or they might use some vocabulary that they don't actually master and that's .hhh how that becomes- just, (.) nonsense

While all FNC teachers did agree that communicativity *should* be the primary criterion used to determine overall essay grade, the interview data suggests that their understanding of communicativity as being interdependent with the other two criteria (cf. Extracts 6–8) means that "Breadth and Accuracy of Language" may have an outsize effect on essay grading. Further confusion in this respect is added by lack of clarity from FMB sensors:

Extract 14 T04:287–91, FNC Scoring Moderation (1)

T04 BECAUSE THE SENSORS WERE REALLY strict on us this time. There were some essays we had no idea (.) where we have gone wrong (.) from this criteria. No idea what had happened.

IR Mmm.

T04 Neither of us could actually (.) discover that.

Extract 15 T05:131–43, FNC Scoring Moderation (2)

IR Yeah, um, is there anything about the, uh, criteria (x) as you see in this ↑ document (.) (x) that differs from (.) how the matriculation board- like, (x) do you (.) view any differences between (.) recommendations by the matriculation board and, hh uh, (. . .) (x) the way you would read this (..) document, if you weren't influenced by the matriculation board? In other words, like- I guess, in- (x) they send you- you're sent this, uh, recommendation about how to ↑ apply (.) the criteria every year, right? No?

T05 We're not. (..) Not really. This is it. We might have seminars, but those are voluntary. So, I mean, there's no (.) we used to receive- but this was even before my time- we used to receive example essays (.) uh, to get some, uh, some kind of- uh, we have example answers (x) in other (x) categories. We used to get those for essays as well. Uh, we don't any more. So it's a bit (.) a shot in the dark, to be honest.

Compared to the IB exam, where subject reports are written within half a year of the exams to clarify how the criteria are applied, the FMB is more opaque about essay scoring compared to IB, with no statistics published on essay scoring and no possibility for teachers to access the sensor rationale for essay moderation (FMB 2020c), whereas marked essays can be purchased by and

mailed to schools in the IB (IBO 2018, 201). Both curricula offer the possibility for essay re-grading. The board also no longer send instructional material on how to apply the criteria, though previous such material may have pre-disposed teachers to prioritizing linguistic competence in producing holistic assessment scores:

Extract 16 *T03:289–96, FNC Scoring Instructions*

T03 the (x) examination board as- examination board do- (. . .) (x) the pedantry (x) comes from them. (x) If you read the things they produce (.) each year in terms of how you're supposed to mark, cause, of course, the teachers mark the exams themselves- when they tell- when they give instructions for teachers for how to mark the exams, you know, if you read these from the years- unless they have changed significantly in fifteen years, °which I would be very surprised if they had,° then they are extremely pedantic (. . .) (x) to a point (x) of ridicule.

While T03's pre-disposition towards FNC essay grading is interesting in terms of how more experienced teachers might grade, the three FNC teachers who do the actual matriculation exam grading began working in English by the time this instructional material was no longer sent out. It is worth noting, nevertheless, that T03's marking of course essays were treated, according to T03, as accurate, if not slightly generous (Extract 12), suggesting that their interpretation of the FNC criteria as prioritizing linguistic competence may still be true.

Overall, the scoring validity of the course essays in both curricula is quite low. Although the pre-DP data suggest that the criteria do measure separate abilities, both IB and FNC essay scores make poor use of the criteria, suggesting that the criteria, as developed for examination, are probably not suitable for course instruction. Furthermore, the available data from FNC essay scores and interviews suggest that matriculation exam criteria do not measure separate abilities and inappropriately emphasize linguistic competence, where emphasis should be placed on, or at least balanced with, sociolinguistic and pragmatic competences. Finally, the lack of feedback between examiners and teachers in FNC compared to IB suggests a lack of a rater training for the FNC matriculation exam essay, which further undermines its scoring validity.

4.2.2 Other Forms of Validity

Assessing other forms of validity, it becomes evident that the FNC curriculum is a victim of its own ambitions (cf. Table 12). While the FNC curriculum seems to demonstrate greater cognitive and criterion-related validity in the closer alignment of its curricular goals with CEFR, the realization of the curriculum in course instruction, examination, and textbook writing (cf., e.g., Table 4) clearly falls short of the broad action-oriented language instruction envision in the curriculum. Opaqueness in exam moderation practices and ambiguity around how to apply the essay criteria (see esp. Extracts 14–15)—alongside discrepancies between the range of written activities

described in the curriculum, on the one hand, compared with the written activities conducted in class and assessed in examination—undermine the consequential and context validity of written assessment in the FNC curriculum. On the other hand, the more limited scope of the IB curriculum, as well as the reliance of IB instruction on exam material for written activities, means that context validity is relatively high, but the reflection of implicitness in language instruction in implicitness of language assessment means that the majority of social contexts for written communication are ignored in the DP curriculum (cf. Table 12), resulting in weak cognitive, consequential, and criterion-related validity. Lastly, some limited remarks can be made about validity of assessment related to test-taker characteristics—namely, a slightly bias in scoring in favor of female students and students with Swedish as their mother tongue.

In the FNC curriculum and textbook, the close correspondence of the range of written production, interaction, and mediation with CEFR suggests strong cognitive validity of assessment in the FNC curriculum, at least at the course level. Tables 4, 10, and 12 indicate how curriculum and course content overlap almost entirely with the activities listed in CEFR for communicative language activities in writing. Thus, there should be strong cognitive validity, in that the types of writing assessed do reflect real-life writing, if the delineation of written communication in CEFR is accepted as an accurate reflection of written communication.

In FNC instruction, however, confusion about how to apply the criteria for the written production task of the matriculation exam, whether as discrete or overlapping criteria (see sub-sec 4.2.1), results in notable confusion about the purpose of written communication, undermining cognitive, context, and consequential validity. This confusion was notably raised when discussing the role of the content of communication in the FNC matriculation exam essay:

Extract 17 T02:78–82, Written Production as Content-free, FNC (1)

T02 it's not a °contents-based° essay.
It's (.) (x) a language (.) test, (.) so.

IR Okay, yeah, [um-]

T02 [(x) SO] WE ALWAYS sa:y that (.) they should not (.) say what they want to, but what they can (.) say °in good English.°

Extract 18 T04:1031–37, Written Production as Content-free, FNC (2)

T04 You can write an essay for the matriculation exam, and if the grammar, and the structure, and the vocabulary, <and the communicativity of it> are perfect, you can write rubbish, really. Or you can write something- (x) you know, isn't necessarily logical at all, but it doesn't matter, because you've written it really really well. What IB does (.) is much more than that. (x) It's about being able (x) to rea::lly think about (.) and articulate (.) something with meaning.

Extract 19 *T05:535–38, Written Production as Content-free, FNC (3)*

T05 I mean, it's supposed to be more (.) about the language, without the contents. We can't really fact check (.) what they're writing about, either, which also is not completely- (x) but you might have different opinions ((laughs)) about (x) whether that's good or bad.

In these extracts, it is striking that all FNC teachers are unanimous in their view that the written production task of the FNC matriculation exam is simply a test of language finesse, an approach to assessment that carries over into written language instruction. Skepticism about the action-oriented nature of the essay criteria (cf. 1Action-oriented language here simply denotes aspects of written production and mediation that focus on meaning instead of communicative linguistic competence. Table 13) was especially pronounced in the informant who taught both IB and FNC:

Extract 20 *T03:434–62, Written Production as Content-free, FNC (4)*

T03 on the one hand, it sounds like the representative of the examination board is (..) is putting forward a progressive idea that (..) that- you know, progressive in the sense of recognizing the importance of language as a communicative act, rather than an academic game of (.) showing that you're better than somebody else- uh, (x) it sounds like a communicative- sort of, a progressive (.) recognition of (x) primacy of communication (x) in ↑ language, .hhh but, at the same time, it sounds like- it's particularly that (.) <it doesn't matter what they write-> you know, this is the whole- (..) this is the whole (..) problem hhh (.) <with the Finnish matriculation board's (.) approach to (..) foreign language teaching> over the years. It doesn't matter what they write. They couldn't care less what they write. They only care (..) whether they're (..) doing it according to the rules that they've decreed as being the only acceptable ones, (..) so I find it, as a question- (x) I find it a little bit hard to respond to, cause I'm not quite sure (..) how to orientate myself towards that- ideological- (x) in terms of the stance being taken. Um, hhh (x) if it's a member of the exam board trying to tell me (.) <that (.) the exam board will not get hung up (..) on the minutiae of how things are expressed,> then I don't believe that for a minute.

IR At least in terms of content-minutiae, not language-minutiae.

T03 Right, in terms of content-minu- (x) yeah, the kind of works, yeah. Yeah, I can-

IR It's a familiar line?

T03 Because content doesn't matter so much. It's not about showing an appreciation of ideas, and understanding of analysis and evaluation. It's just about (..) the mechanical construction (x) of sentences. It's, like, (. . .) programming artificially intelligent (.) language, kind of stuff, (.) you know. (..) No, that's unfair on the artificial intelligence community, I think.

As all informants reported that the content of communication is not an important aspect of written communication, as FNC instruction is organized such that teachers are the primary markers of the matriculation exam, and as there are notable discrepancies between the range of activities given in the FNC curriculum and the range of activities evidenced in course and matriculation exam assessment (see Table 12), the results here are alarming for the cognitive, context, and consequential validity of assessment. Cognitive validity of FNC written assessment is undermined in that assessment of writing seems to focus only on language finesse, what CEFR refers to as

communicative linguistic competence, rather than sociolinguistic, pragmatic, and task-type-related content considerations. This departure from explicit action-oriented pedagogy envisioned in the FNC curriculum (Juurakko-Paavola 2019; Opetushallitus 2019, 174–79) suggests weak context validity, as the skills assessed do not match the skills that the curriculum is intended to develop. Given that this discrepancy arises in both the matriculation exams and course instruction in four out of the five types of written production activities that appear in the FNC curriculum and CEFR but not in examination or course instruction (see Table 12), it is highly likely that course assessment was heavily influenced by the high-stakes nature of the FNC matriculation exam, suggesting weak consequential validity. That is, the absence of certain communicative language activities and strategies, as well as the assessment thereof, in the matriculation exams has shifted the focus of course instruction and assessment to only the types of written activities and assessment prioritized in the matriculation exam.

The implicit nature of language teaching in the IB, on the other hand, and the curriculum's focus on academic writing allows for greater cognitive and context validity but significantly weaker criterion-related validity. Like assessment in FNC, consequential validity is also weakened by the focus of course assessment on the types of writing in the high-stakes final IB exam. Given the high degree of correspondence between curriculum, course assessment, and exam assessment of communicative language activities in writing (cf. Table 12), context validity is high in the IB. Context validity seems also to be bolstered by the annual publication of subject reports to guide teachers on aligning course assessment to exam assessment (see sub-sec 2.3.2). Unfortunately, I neglected to include a question related to this point in the interview guide, so the qualitative data do not evidence whether the subject reports actually have contributed to context validity in IB written assessment. On the other hand, the interview data did suggest that cognitive validity was high in IB assessment, as there was greater focus on language action instead of language mastery, as seen in the FNC data:

Extract 21 *T01:305–11, Written Production as Content-rich, IB (1)*

T01 Yeah, cause I think it (x) I think it's more important, the content, like, the actual analysis (.) and understanding (..) is more (..) important.

IR And why do you think that's more important?

T01 Um (. . .) well, because, if- I think if you can communicate to the point where someone can understand what you're trying to ↑ say, then that's (. . .) ↑ good (..) good ↑ enough. But then if you haven't got anything to ↓ say, that's not good.

Extract 22 *T03:305–11, Written Production as Content-rich, IB (2)*

T03 The only thing I do sometimes say about language is I do sometimes say, you know, (x) you might want to proofread it a bit more carefully, cause (..) I sometimes suspect students are (..) hhh (x) are letting a lot of errors through that they could quite easily (.) sift out, if they just went through it. (x) Other

> than that, I don't- so (x) so I suppose, for me, the balance (..) is (.) much more about content (x) than about (..) grammar. (x) I'm much more- (.) I'm much more interesting in developing students' critical reading ability (..) and their ability to (.) uh, express (.) complex ideas succinctly- (..) succinctly and efficiently and effectively- (..) than I am about (..) their ability to (..) put a comma in the right place.

Both IB teachers thus agreed that formative and summative assessment of writing in the IB should focus on the action orientation of language rather than language mastery.

While these data suggests strong internal validity in the IB, there are alarmingly few task types evidence in IB course assessment and examination, such that external comparisons with the FNC curriculum and CEFR suggest low criterion-related validity. The paucity of communicative language activities present in DP English is especially alarming for pre-DP and DP instruction of English in Finland, where a majority of students are EFL users, at least at the school studied (cf. Table 9). Given that teachers in the IB are aware that the majority of their students are EFL users (cf. T01:691–702; T04:735–40), the influence of IB exam matriculation on course assessment—as opposed to, for instance, influence from their FNC colleagues—suggests weak consequential validity, as course assessment is catered towards the high-stakes final examination of the curriculum.

Finally, as for validity related to test-taker characteristics, gender and language were tested for correlation and differences in mean values of the final essays written in each first-year English course. Bivariate correlation was calculated between gender and FNC exam equivalent achievement band results (1–7) for the essays, a value calculated based on the proportion of the scores and rounded to the nearest values available in FNC examination to make the scores comparable (FMB 2017, 16). Pearson's *r* was calculated for gender and the three final scores in the cohort as a whole and between IB and FNC students. While all cases yielded a positive *r* value, all but one result was statistically insignificant—namely, the only statistically significant result found was the limited correlation observed between gender the final essay score of IB students in their second course ($r = 0.443$, $p < 0.01$, $n = 45$). These results indicate a slightly tendency overall for female students at the school studied to receive higher grades than male students.

Given the diversity of languages spoken by students at the school studied (cf. Table 9), the effect of language on essay scoring was more difficult to assess. To facilitate comparison, students were divided into those whose mother tongues are registered in the population registry as Swedish, Finnish, English, and other (on the unreliability of the population registry for student's mother tongue, see sub-sec 3.2.2):

CURRICULUM	L1		ENA1	ENA2	ENA3
FNC	Swedish	Mean	6.32	6.30	6.34
		n	101	101	101

		Std. Deviation	.747	.729	.739
	Finnish	Mean	6.31	6.08	6.00
		n	13	13	13
		Std. Deviation	.630	.760	.577
	Total	Mean	6.32	6.27	6.30
		N	114	114	114
		Std. Deviation	.733	.732	.728
Pre-DP	Swedish	Mean	6.20	6.13	6.53
		n	15	15	15
		Std. Deviation	.775	.516	.516
	Finnish	Mean	6.50	5.33	6.17
		n	12	12	12
		Std. Deviation	.522	.888	.718
	English	Mean	6.43	5.57	6.29
		n	7	7	7
		Std. Deviation	.787	1.272	1.113
	Other	Mean	6.27	5.18	6.00
		n	11	11	11
		Std. Deviation	.786	1.168	1.095
	Total	Mean	6.33	5.60	6.27
		N	45	45	45
		Std. Deviation	.707	.986	.837

Table 18: Mean Scores of Essays (Achievement Band), by Language Group and Curriculum

As students with Swedish as their mother tongue still represented the overwhelming majority (72.96%) of students in the sample and the distributions of the data sets are non-normal, a Mann-Whitney U test was conducted comparing students with Swedish as their mother tongue versus those with other languages as their mother tongue. Based on these results, statistical significance was found in the third course—$U = 463$, $p < .05$—of FNC English and the second course—$U = 83.5$, $p < .001$—of IB English. Comparing these results with the data presented in Table 18, there are at least indications that students with Swedish as their mother tongue tend, on average, to receive statistically significantly higher essay scores. Interestingly, the only time ENL users on average scored higher than students with Swedish as their mother tongue was in the first course; a Mann-Whitney U test comparing students with English and those with Swedish as their mother tongues was conducted, however, indicating no statistically significant difference between the two distributions—$U = 31.5$, $p = 0.122$. This result is inconclusive, however, given the significant differences in sample size between students with English and those with Swedish as their mother tongue.

Altogether, these results indicate relatively strong validity related to test-taker characteristics. However, the slight tendency for female students to score higher than male students,

especially in pre-DP assessment, and for students with Swedish as their mother tongue to score higher than students with other languages as their mother tongue suggest that differentiation, scorer training, adjustment of examination procedure, or any combination of the three is needed to mitigate the effects of test-taker characteristics on written assessment.

5 Conclusion

5.1 Summary of Results

In section 2.4, I gave the following hypotheses:

- *Hypothesis 1*: FNC reflects a larger shift towards action-oriented language production compared to IBDP, (1a) except in more complex forms of production and mediation tasks compared to DP, (1b) in more varied forms of interaction tasks, as a result of backwash, (1c) in its reliance of FonFS instruction.
- *Hypothesis 2*: The FNC curriculum exhibits lower cognitive, context, scoring, and criterion-related validity, (2a) with both curricula scoring low for consequential validity, and (2b) a larger gender performance gap in the IB.
- *Hypothesis 3*: Best practices emphasizing the instruction and assessment of action-oriented language draw attention to language as a tool of communication, the noticing of language, and a combined FonF/FonFS approach to language acquisition.

The results of this thesis indicate mixed results, confirming and rejecting various of these hypotheses.

Hypothesis 1 does largely seem to be confirmed from the data. In particular, the results gathered in Table 12 indicate that action-oriented language is better represented in FNC curriculum, course instruction and assessment, and examination than in IBDP. Hypothesis 1a received some support based on the results gathered in Table 4 (cf. Table 12) but seemed to be an oversimplification. While examination in the FNC curriculum was found to focus on non-academic forms of writing, more complex forms of written production and mediation activities were found in course instruction and assessment, in addition to the much wider breadth of activities found in the latter, confirming hypothesis 1b. On the other hand, hypothesis 1c did receive some support from the data presented here, especially as indicated in Extracts 17–20, where an FonFS approach seemed to be a common theme in reasons why written assessment deviated from action-oriented language assessment.

Hypothesis 2 was the most mixed. Overall, the FNC curriculum had higher cognitive and criterion-related validity in theory (sec 4.1), though, in practice, cognitive, context, and criterion-related validity were undermined by confusion around how to apply the written production criteria (including whether to treat the criteria discretely, opaqueness of essay score moderation, and resistance to changing standards of assessment) with language mastery emerging as a focus for course instruction and assessment over the action-oriented approach of treating language as a tool. On the other hand, while IB assessment evidenced strong cognitive, context, and scoring validity,

especially in the discrete application of CR assessment (cf. Table 15), the strong context and cognitive validity were greatly limited by the weak criterion-related validity in the educational line. That is, the close correspondence of curriculum to assessment and of written task to real-life equivalence were invalidated by the weak correspondence of either with the much wider range of communicative language activities in writing recommended in CEFR, a particularly concerning weakness for IB instruction in Finland, where students are largely EFL users. It is thus difficult to say whether there is a less valid education line, as both raise serous concerns.

In both educational lines, the influence of high-stakes examination seemed to explain discrepancies between curriculum and assessment on the FNC side and between the demands of the curriculum and the needs of students demographic on the IB side, confirming hypothesis 2a. Hypothesis 2b was largely rejected, as limited correlation was found between gender and essay scores in one of six correlations, the remainder of which did not reach statistical significance, though all correlations did point uniformly to higher results among female students. Additionally, the results pointed tentatively to higher results in written assessment among students with Swedish as their mother tongue.

Lastly, hypothesis 3 was not able to be explored in the limited space of this thesis. Nevertheless, the data were gathered in the interviews conducted and can be found in the transcripts, the file of which can be found from the link in Appendix G. The relevant data are coded "atl_metacognition," the first three letters standing for "approaches to learning." Those data point uniformly to a recognition on the part of each informant of importance of metacognition in EFL instruction and assessment.

5.2 Limitations and Further Research

Due to the limited scope and space of this thesis, not all the data gathered for this thesis could be explored in depth. Analysis of the curriculum, for instance, was done largely in summaries rather than systematically. CAQDA was conducted only with the interview data, given that I had to complete the interviews, transcription, coding, and data analysis on my own. Quantitative analysis has limited wider implications due to the methodological focus of this thesis on approaching the topic using a case study and the intended purpose of the thesis as inform the direction of further action research. Thus, the results will be especially helpful for developing instruction and assessment at the school but may find limited utility outside the context of the school.

Future research should look to compare these results with other schools offering IB and FNC lines, considering the substantial lack of such comparative research in Finland outside of those

conducted in bachelor's and master's theses. Furthermore, the alarmingly low cognitive, context, scoring, and consequential validity of instruction and assessment in the FNC curriculum, on the one hand, and the alarmingly low criterion-related and consequential validity of instruction and assessment in the IB curriculum at Finnish schools both need to be addressed. Especially from the perspective of CEFR, both curricula are still far behind in implementing action-oriented language instruction and assessment at both course and exam levels. Further studies should also investigate the extent to which principles of CEFR are supported in textbooks used in Finnish, both in theory and in practice. While the FNC curriculum evidenced a wider embrace of action-oriented language pedagogy, its realization in course and exam assessment is still lagging behind. Lastly, not all of the interview data were assessed here and can still be further used, for instance, in the comparative evaluation of best practices of EFL teachers in IB and FNC.

As for action research, this project now moves on to the process of implementing these results when combining FNC and pre-DP EFL instruction. Due to the COVID-19 pandemic, teacher meetings have been limited, such that I was not able to introduce my research to my colleagues as of the time this thesis was submitted. Most likely, then, development of a common pre-DP and FNC curriculum can only be completed for academic year 2022–2023.

5.3 Implications for Development of Course and Exam Assessment

The results of this master's thesis, which have benefitted from comparative analysis, indicate that the following areas of course and exam assessment can be improved. For IB assessment in Finland, the limited range of written production and interaction needs to be addressed at the school level, both in DP and pre-DP instruction and assessment. Within the DP curriculum, for instance, different forms of communicative language activities in writing can be completed in the Learner Portfolio (IBO 2019b, 25–26). Given that the Learner Portfolio is not directly assessed by the IB, teachers and schools should invest resources to ensure adequate formative assessment here, given the lack of communicative language activities in the mandatory parts of the curriculum and the tendency for the demands of high-stakes evaluation to crowd out parts of the curriculum that are not assessed.

As for FNC assessment, transparency is desperately needed from the FMB in general and the sensors in particular. The language of the criteria is especially vague, with no clarity on whether the criteria should be treated discretely or how they should be weighted, while the interview data suggested that there is a tendency for assessment practices to continue regardless of changes in criteria description, especially in the focus on linguistic competence in favor of sociolinguistic and

pragmatic competences as well as proficiency in communicative language activities and strategies. Given that a wider range of written production, interaction, and mediation are explicitly required by the FNC curriculum, there is also a greater need in FNC examination to evaluate which activities are actually used in examination and which have been ignored, as these were seen to have a profoundly negative washback effect on course instruction and assessment. While continuum CR assessment of written production in the matriculation exam is invaluable for improving student writing, its use for mastery NR assessment is bewildering, considering the broad range of EFL learning in Finland and its incongruity with continuum CR assessment, and should be re-evaluated by the FMB.

Finally, while this thesis was not able to formally address the question of the comparability of the two curricula evaluated for combined first-year high school instruction, its use of CEFR as a comparative tool for assessing action-oriented language instruction and assessment in both curricula suggest that pre-DP instruction in Finland can and should be more closely aligned with FNC instruction, particularly in light of the closer alignment of the FNC curriculum with CEFR. The pre-DP student population of Finland is poorly served by the legislative ambiguity of their curriculum and need to be supported by a strong basis in action-oriented language instruction, given that the IB falls much shorter of alignment with CEFR than does the FNC curriculum, particularly for a predominately EFL student body. Reference material, like that produced in Appendix A, should also be consulted by teachers and examiners to ensure adherence to CEFR guidelines.

Bibliography

Abbreviations

CEFR	*Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Council of Europe 2001)
IBO	International Baccalaureate Organization
FMB	Finnish Matriculation Board
NARIC	National Academic Recognition Information Centres
OSF	Official Statistics Finland

References

Akcin, Dogan Can. 2019. “Efficacy of Explicit and Implicit Instructions on the Acquisition and Production of the English Passive.” London: University College London.

Andtfolk, Martina, Camilla Hannuksela, and Harriet Lindroth. 2016. *New Profiles 2*. Schildts & Söderströms.

Carrió-Pastor, María Luisa, and Inmaculada Tamarit Vallés. 2015. “A Comparative Study of the Influence of the Mother Tongue in LSP and CLIL.” *Procedia Social and Behavioral Sciences* 178 (1): 38–42.

Chandra, Yanto, and Liang Shang. 2017. “An RQDA-Based Constructivist Methodology for Qualitative Research.” *Qualitative Market Research: An International Journal* 20 (1): 90–112. https://doi.org/10.1108/QMR-02-2016-0014.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Creswell, John W. 2013. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*.

Dalton-Puffer, Christiane. 2011. “Content-and-Language Integrated Learning: From Practice to Principles?” *Annual Review of Applied Linguistics* 31: 182–204.

Dweck, Carol S. 2006. *Mindset: The New Psychology of Success*. New York: Random House.

Ellis, Rod. 2009. *Task-Based Language Learning and Teaching*. 7. print. Oxford Applied Linguistics. Oxford: Oxford Univ. Press.

———. 2012. *Form-Focused Instruction and Second Language Learning*. Oxford: Wiley-Blackwell.

———. 2015. “Form-Focused Instruction and the Measurement of Implicit and Explicit L2 Knowledge.” In *Implicit and Explicit Learning of Languages*, edited by Patrick Rebuschat, 48:417–41. Studies in Bilingualism. Amsterdam: John Benjamins.

EU and Council of the EU. 2016. *General Data Protection Regulation*. *2016/679*.

Ferman, Irit. 2004. “The Washback of an EFL National Oral Matriculation Test to Teaching and Learning.” In *Washback in Language Testing: Research Contexts and Methods*, edited by Liying Cheng and Yoshinori Watanabe, 191–210. Mahwah, N.J: Lawrence Erlbaum Associates.

Figueras, Neus. 2012. “The Impact of the CEFR.” *ELT Journal* 66 (4): 477–85.

FMB. 2017. “Toisen kotimaisen kielen ja vieraiden kielten digitaalisten kokeiden määräykset.” Helsinki: FMB. https://www.ylioppilastutkinto.fi/images/sivuston_tiedostot/Ohjeet/Koekohtaiset/sahkoiset_kielikokeet_maaraykset_30.11.2017_fi.pdf.

———. 2019. “Englanti (Pitkä Oppimäärä, 2019 Kevät).” Yle. May 24, 2019. http://yle.fi/plus/abitreenit/2019/kevat/EA-fi/index.html#q16.

———. 2020a. “Pisterajat.” Ylioppilastutkintolautakunta. 2020. https://www.ylioppilastutkinto.fi/ylioppilastutkinto/pisterajat.

———. 2020b. “Tilastotaulukot.” Ylioppilastutkintolautakunta. 2020. https://www.ylioppilastutkinto.fi/tietopalvelut/tilastot/tilastotaulukot.
———. 2020c. “Tutkintotulokset Ja Koesuoritukset.” Ylioppilastutkintolautakunta. 2020. https://www.ylioppilastutkinto.fi/ylioppilastutkinto/tulokset-ja-koesuoritukset.
———. 2020d. “Tulokset Ja Koesuoritukset.” May 7, 2020. https://www.ylioppilastutkinto.fi/ylioppilastutkinto/tulokset-ja-koesuoritukset.
González-Lloret, Marta. 2019. “Task-Based Language Teaching and L2 Pragmatics.” In *The Routledge Handbook of Second Language Acquisition and Pragmatics*, edited by Naoko Taguchi, 338–52. New York: Routledge.
Goo, Jaemyung, Gisela Granena, Yucel Yilmaz, and Miguel Novella. 2015. “Implicit and Explicit Instruction in L2 Learning.” In *Implicit and Explicit Learning of Languages*, edited by Patrick Rebuschat, 48:443–82. Studies in Bilingualism. John Benjamins Amsterdam.
Hansen, Petteri. 2006. “Selvitys Suomen IB-ylioppilaiden koetuloksista ja IB-tutkinnolla korkeakouluihin sijoittumisesta vuosina 2001–2004.” Helsinki: Opetusministeriö.
Hauta-Aho, Sanna. 2013. “IB Upper Secondary School and National Upper Secondary School Students’ Attitudes towards English Oral Skills: A Comparative Study.” Master’s Thesis, Jyväskylä: University of Jyväskylä.
House, Juliane. 2014. “English as a Global Lingua Franca: A Threat to Multilingual Communication and Translation?” *Language Teaching* 47 (3): 363–76. https://doi.org/10.1017/S0261444812000043.
Hurd, Emily. 2017. “Translation—The Fifth Language Skill? A Comparison of the Role of Translation in Finnish Lukio and the International Baccalaureate Diploma Programme.” Turku: University of Turku.
IBO. (2011) 2013. “Language A: Language and Literature Guide.” IBO.
———. 2017. “Grade Descriptors: For Use from December 2017.” Cardiff: IBO.
———. 2018. “Assessment Principles and Practices—Quality Assessments in a Digital Age.” Cardiff: IBO.
———. 2019a. “English A: Language and Literature (Subject Report, May 2019).” Cardiff: IBO.
———. 2019b. “Language A: Language and Literature Guide.” IBO.
———. (2018) 2019. “Language B Guide.” IBO.
———. 2020a. “Country Profile: Finland.” International Baccalaureate. 2020. https://www.ibo.org/about-the-ib/the-ib-by-country/f/finland/.
———. 2020b. “The IB Diploma Programme Provisional Statistical Bulletin: November 2019 Examination Session.” Pearson.
———. 2020c. “May 2020 Examinations Will No Longer Be Held.” International Baccalaureate. March 23, 2020. https://www.ibo.org/news/news-about-the-ib/may-2020-examinations-will-no-longer-be-held/.
Ivankova, Nataliya V., and Jennifer L. Greer. 2015. “Mixed Methods Research and Analysis.” In *Research Methods in Applied Linguistics: A Practical Resource*, edited by Brian Paltridge and Aek Phakiti, 63–81. London: Bloomsbury.
Junger, Solveig. 1999. “The International Baccalaureate section at Vasa övningsskola.” In *Content and Language Integrated Learning: Teachers’ and Teacher Educators’ Experiences of English Medium Teaching*, edited by Kaj Sjöholm and Mikaela Björklund. Vasa: Faculty of Education, Åbo Akademi University.
Juurakko-Paavola, Taina. 2019. “Relating Finnish Matriculation Examination Grades to the CEFR.” In *Developments in Language Education: A Memorial Volume in Honour of Sauli Takala*, edited by Ari Huhta, Gudrun Erickson, and Neus Figueras, 147–51. Jyväskylä: EALTA and University of Jyväskylärelating.

Kainulainen, Tiina. 2006. “Understanding Idioms: A Comparison of Finnish Third Grade Students of National Senior Secondary School and IB Diploma Programme.” Master’s Thesis, Jyväskylä: University of Jyväskylä.

Kansainvälisiä koulutustarpeita käsittelevä työryhmä. 2007. “Kansainväliset opetustarpeet: IB-tutkinnon järjestäminen Suomessa.” Opetusministeriön työryhmämuistioita ja selvityksiä. Opetusministeriö.

Karusigarira, Anni. 2016. “Error Analysis of the Written English of Finnish High School Students in National and IB Programs.” Master’s Thesis, Helsinki: University of Helsinki.

Knoch, Ute, and Carol A. Chapelle. 2018. “Validation of Rating Processes within an Argument-Based Framework.” *Language Testing* 35 (4): 477–99. https://doi.org/10.1177/0265532217710049.

Kolehmainen, Tiina. 2014. “Perceptions, Attitudes and Uses of English: A Comparative Study of Finnish Students in the International Baccalaureate Diploma Programme and the Finnish Upper Secondary School System.” Master’s Thesis, Joensuu: University of Eastern Finland.

Kovanen, Heidi. 2011. “How CLIL-Classroom Students See Themselves as Learners of English.” Master’s Thesis, Jyväskylä: University of Jyväskylä.

Lee, Sang-Ki, and Hung-Tzu Huang. 2008. “Visual Input Enhancement and Grammar Learning: A Meta-Analytic Review.” *Studies in Second Language Acquisition,* 307–31.

Li, Shuai. 2019. “Cognitive Approaches in L2 Pragmatics Research.” In *The Routledge Handbook of Second Language Acquisition and Pragmatics,* edited by Naoko Taguchi, 113–27. New York: Routledge.

Lindroth, Harriet, Camilla Hannuksela, and Martina Rosenback. 2016. *New Profiles 1*. Schildts & Söderströms.

Loes, Chad, Ernest Pascarella, and Paul Umbach. 2012. “Effects of Diversity Experiences on Critical Thinking Skills: Who Benefits?” *The Journal of Higher Education* 83 (1): 1–25. https://doi.org/10.1353/jhe.2012.0001.

Long, Michael H. 1991. “Focus on Form: A Design Feature in Language Teaching Methodology.” In *Foreign Language Research in Cross-Cultural Perspective*, edited by Kees de Bot, Ralph Ginsberg, and Claire Kramsch, 39–52. Amsterdam: John Benjamins.

———. 1996. “The Role of the Linguistic Environment in Second Lanugage Acquisition.” In *Handbook of Second Language Acquisition,* edited by William C. Ritchie and Tej K. Bhatia, 413–68. San Diego: Academic Press.

Määttä, Liisa. 2014. “‘The World Has Gotten Smaller’: Third-Year IB High School Students’ Perceptions on the International Baccalaureate Diploma Programme: Internationalisation, English Medium Instruction and Career Choice.” Master’s Thesis, Jyväskylä: University of Jyväskylä.

Martikainen, Minna. 2020. “Inkeriläisten paluumuutto ja kielitutkinto sen osana.” *Kieliverkosto* 11 (1): n.p.

Maurette, Marie-Thérèse. 1948. “Educational Techniques for Peace: Do They Exist?” UNESCO.

McMillan, James H., and Jessica Hearn. 2008. “Student Self-Assessment: The Key to Stronger Student Motivation and Higher Achievement.” *Educational Horizons* 87 (1): 40–49.

Nikku, Juha. 2019. “Etelä-Karjalan IB-lukion imago, maine ja mielikuva.” Bachelor’s Thesis, Lapeenranta: Saimaa University of Applied Sciences.

Norris, John M., and Lourdes Ortega. 2000. “Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-Analysis.” *Language Learning* 50 (3): 417–528. https://doi.org/10.1111/0023-8333.00136.

North, Brian, Tim Goodier, and Enrica Piccardo. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Companion Volume with New Descriptors)*.

Nylund, Anna-Liisa. 2017. “Taideproduktioiden merkitys lukio-opetuksessa: Lahden kaupungin lukioiden opiskelijoiden kokemuksia.” Master’s Thesis, Helsinki: Humak University of Applied Sciences.

OECD. 2015. *Education Policy Outlook*. OECD Publishing. http://www.oecd-ilibrary.org/education/education-policy-outlook-2015_9789264225442-en.

Opetushallitus. 2015. “Lukion opetussunnitelman perusteet.” Grano.

———. 2019. “Lukion opetussunnitelman perusteet.” Grano.

OSF. 2020. “Upper Secondary General School Education.” Statistics Finland. June 16, 2020. https://www.stat.fi/til/lop/index_en.html.

Rautio, Marjatta, and Reeta Niemonen. 2020. “Katso, millä keskiarvolla Suomen eri lukioihin pääsi – Jani Ollila aloittaa erälukion Enontekiöllä, Juuso Laine matikkalukion Espoossa.” *Yle*, June 17, 2020. https://yle.fi/uutiset/3-11395333.

Richards, Jack C., and Theodore S. Rodgers. 2001. *Approaches and Methods in Language Teaching*. 2nd edition. Cambridge Language Teaching Library. Cambridge: Cambridge Univ. Press.

Richards, Keith. 2003. *Qualitative Inquiry in TESOL*. New York: Palgrave Macmillan.

Rinta-Kiikka, Suvi, Tapani Yrjölä, and Eeva Alho. 2018. “Talous, arvot ja alueellinen sosiaalinen pääoma.” Helsinki: Pellervon taloustutkimus.

Rosenback, Martina, Harriet Lindroth, Camilla Hannuksela, and Sarah Mattila. 2017. *New Profiles 3*. Schildts & Söderströms.

Ryshina-Pankova, Maryanna. 2019. “Systemic Functional Linguistics and L2 Pragmatics.” In *The Routledge Handbook of Second Language Acquisition and Pragmatics*, edited by Naoko Taguchi, 255–71. New York: Routledge.

Schmidt, Richard W. 1990. “The Role of Consciousness in Second Language Learning.” *Applied Linguistics* 11 (2): 129–58.

Schwarzenthal, Miriam, Maja K. Schachner, Linda P. Juang, and Fons J. R. van de Vijver. 2020. “Reaping the Benefits of Cultural Diversity: Classroom Cultural Diversity Climate and Students’ Intercultural Competence.” *European Journal of Social Psychology* 50 (2): 323–46. https://doi.org/10.1002/ejsp.2617.

Shaw, Stuart D., and Cyril J. Weir. 2007. *Examining Writing: Research and Practice in Assessing Second Language Writing*. Studies in Language Testing 26. Cambridge: Cambridge University Press.

Shin, Hye Won. 2010. “Another Look at Norris and Ortega (2000).” *Studies in Applied Linguistics and TESOL* 10 (1).

Spada, Nina, and Patsy Lightbown. 2013. “Instructed Second Language Acquisition.” In *The Routledge Encyclopedia of Second Language Acquisition*, edited by Peter Jake Robinson, 319–27. New York: Routledge.

Spada, Nina, and Yasuyo Tomita. 2010. “Interactions Between Type of Instruction and Type of Language Feature: A Meta-Analysis.” *Language Learning* 60 (2): 263–308. https://doi.org/10.1111/j.1467-9922.2010.00562.x.

Tamminen, Susanna. 2005. “IB-opinto-ohjauksen kehittäminen Ressun lukiossa.” *Ohjauksen kehittämishankkeita ja käytänteitä*. Jyväskylä: University of Jyväskylä.

Tateyama, Yumiko. 2019. “Pragmatics in a Language Classroom.” In *The Routledge Handbook of Second Language Acquisition and Pragmatics*, edited by Naoko Taguchi, 400–412. New York: Routledge.

Tirri, Kirsi. 2014. “The Last 40 Years in Finnish Teacher Education.” *Journal of Education for Teachering* 40 (5): 600–609.

Truscott, John. 2007. “The Effect of Error Correction on Learners’ Ability to Write Accurately.” *Journal of Second Language Writing* 16 (4): 255–72.

UK NARIC. 2016. “Benchmarking Selected IB Diploma Programme Language Courses to the Common European Framework of Reference for Languages.” IBO. https://www.ibo.org/globalassets/publications/ib-research/dp/ib-dp-cefr-benchmarking-report-en.pdf.

Wallace, Michael J. 2006. *Action Research for Language Teachers*. Cambridge: Cambridge University Press.

Weir, Cyril J. 2005. *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Wells, Amy Stuart, Lauren Fox, and Diana Cordova-Cobo. 2016. “How Racially Diverse Schools and Classrooms Can Benefit All Students.” *The Education Digest* 82 (1): 17.

World Bank. 2020. “World Bank Open Data.” 2020. https://data.worldbank.org/.

Appendix A CEFR (2018) Reference Level Descriptors, Writing (B2.1)

<table>
<tr><th colspan="5">COMMUNICATIVE LANGUAGE ACTIVITIES AND STRATEGIES</th></tr>
<tr><td colspan="3">Written Production
Overall[1]
• Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources.</td><td colspan="2">Interaction
Overall[1]
• Can express news and views effectively in writing, and relate to those of others.</td></tr>
<tr><td colspan="2">Creative writing
• Can write straightforward, detailed descriptions on a range of familiar subjects within his/her field of interest.
• Can write accounts of experiences, describing feelings and reactions in simple connected text.
• Can write a description of an event, a recent trip—real or imagined.
• Can narrate a story.</td><td>Written report and essays
• Can write an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options.
• Can synthesise information and arguments from a number of sources.</td><td>Correspondence
• Can write letters conveying degrees of emotion and highlighting the personal significance of events and experiences and commenting on the correspondent's news and views.
• Can use formality and conventions appropriate to the context when writing personal and professional letters and emails.
• Can write formal emails/letters of invitation, thanks or apology with appropriate register and conventions.
• Can write non-routine professional letters, using appropriate structure and conventions, provided these are restricted to matters of fact.
• Can obtain, by letter or e-mail, information required for a particular purpose, collate it and forward it by mail to other people.</td><td>Notes, messages, and forms[1]
• Can take or leave complex personal or professional messages, provided he/she can ask clarification or elaboration if necessary</td></tr>
<tr><td colspan="3">Production strategies</td><td colspan="2">Interaction strategies</td></tr>
<tr><td rowspan="2">Planning
• Can plan what is to be said and the means to say it, considering the effect on the recipient(s).</td><td rowspan="2">Compensating
• Can address most communication problems by using circumlocutions, or by avoiding difficult expressions.</td><td rowspan="2">Monitoring and repair
• Can correct slips and errors if he/she becomes conscious of them or if they have led to misunderstandings.
• Can make a note of ‘favourite mistakes’ and consciously monitor speech for it/them.</td><td>Taking the floor (turntaking)[1]
• Can intervene appropriately in discussion, exploiting appropriate language to do so.
• Can initiate, maintain and end discourse appropriately with effective turn taking.
• Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly.
• Can use stock phrases (e.g. ‘That's a difficult question to answer’) to gain time and keep the turn whilst formulating what to say.</td><td>Cooperating
• Can help the discussion along on familiar ground, confirming comprehension, inviting others in, etc.
• Can summarise the point reached at a particular stage in a discussion and propose the next steps.</td><td>Asking for clarification
• Can, in informal conversation (with friends), ask for explanation or clarification to ensure he/she understands complex, abstract ideas.
• Can formulate follow-up questions to a member of a group to clarify an issue that is implicit or poorly articulated.</td></tr>
<tr><td colspan="3">Online interaction</td></tr>
<tr><td colspan="2">Online conversation and discussion
• Can participate actively in an online discussion, stating and responding to opinions on topics of interest at some length, provided contributors avoid unusual or complex language and allow time for responses.
• Can engage in online exchanges between several participants, effectively linking his/her contributions to previous ones in the thread, provided a moderator helps manage the discussion.
• Can recognise misunderstandings and disagreements that arise in an online interaction and can deal with them, provided that the interlocutor(s) are willing to cooperate.</td><td>Goal-oriented online transaction and collaboration
• Can collaborate online with a group that is working on a project, justifying proposals, seeking clarification and playing a supportive role in order to accomplish shared tasks.</td></tr>
</table>

Mediation of a text

Overall

- Can work collaboratively with people from different backgrounds, creating a positive atmosphere by giving support, asking questions to identify common goals, comparing options for how to achieve them and explaining suggestions for what to do next.
- Can further develop other people's ideas, pose questions that invite reactions from different perspectives and propose a solution or next steps. Can convey detailed information and arguments reliably, e.g. the significant point(s) contained in complex but well-structured texts within his/her fields of professional, academic and personal interest.

Relaying specific information in writing[2]	*Explaining data in writing (e.g. in graphs, diagrams, charts etc.)*[1, 2]	*Processing text in writing*	*Translating a written text in writing*	*Note-taking (lectures, seminars, meetings etc.)*[1]	*Expressing a personal response to creative texts (including literature)*[1]	*Analysis and criticism of creative texts (including literature)*[1]
• Can relay in a written report (in Language B) relevant decisions that were taken in a meeting (in Language A). • Can relay in writing the significant point(s) contained in formal correspondence (in Language A).	• Can interpret and present reliably in writing (in Language B) detailed information from diagrams and visually organised data in his fields of interest (with text in Language A).	• Can summarise in writing (in Language B) the main content of complex spoken and written texts (in Language A) on subjects related to his/her fields of interest and specialization.	• Can produce translations into (Language B), which closely follow the sentence and paragraph structure of the original text in (Language A), conveying the main points of the source text accurately, though the translation may read awkwardly.	• Can understand a clearly structured lecture on a familiar subject, and can take notes on points which strike him/her as important, even though he/she tends to concentrate on the words themselves and therefore to miss some information. • Can make accurate notes in meetings and seminars on most matters likely to arise within his/her field of interest	• Can give a clear presentation of his/her reactions to a work, developing his/her ideas and supporting them with examples and arguments. • Can describe his/her emotional response to a work and elaborate on the way in which it has evoked this response. • Can express in some detail his/her reactions to the form of expression, style and content of a work, explaining what he/she appreciated and why.	• Can compare two works, considering themes, characters and scenes, exploring similarities and contrasts and explaining the relevance of the connections between them. • Can give a reasoned opinion about a work, showing awareness of the thematic, structural and formal features and referring to the opinions and arguments of others. • Can evaluate the way the work encourages identification with characters, giving examples. • Can describe the way in which different works differ in their treatment of the same theme.

COMMUNICATIVE LANGUAGE COMPETENCES

Linguistic					**Sociolinguistic**	**Pragmatic**				
Linguistic Range	*Vocabulary Range*	*Grammatical accuracy*	*Vocabulary control*[1]	*Orthographic Control*[1]	*Sociolinguistic appropriateness*	*Flexibility*	*Turntaking*[1]	*Thematic development*	*Coherence and cohesion*	*Propositional precision*
• Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so.	• Has a good range of vocabulary for matters connected to his/her field and most general topics. • Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution. • Can produce the appropriate collocations of many words in most contexts fairly systematically. • Can understand and use much of the specialist vocabulary of his/her field but has problems with specialist terminology outside of it.	• Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding. • Has a good command of simple language structures and some complex grammatical forms, although he/she tends to use complex structures rigidly with some inaccuracy.	• Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.	• Can produce clearly intelligible continuous writing, which follows standard layout and paragraphing conventions. • Spelling and punctuation are reasonably accurate but may show signs of mother tongue influence.	• Can adjust his/her expression to make some distinction between formal and informal registers but may not always do so appropriately. • Can express him/herself appropriately in situations and avoid crass errors of formulation. • Can sustain relationships with speakers of the target language without unintentionally amusing or irritating them or requiring them to behave other than they would with another proficient speaker.	• Can adjust to the changes of direction, style and emphasis normally found in conversation. • Can vary formulation of what he/she wants to say. • Can reformulate an idea to emphasise or explain a point.	• Can intervene appropriately in discussion, exploiting appropriate language to do so. • Can initiate, maintain and end discourse appropriately with effective turn taking. • Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. • Can use stock phrases (e.g. 'That's a difficult question to answer') to gain time and keep the turn whilst formulating what to say.	• Can follow the conventional structure of the communicative task concerned, when communicating his/her ideas. • Can develop a clear description or narrative, expanding and supporting his/her main points with relevant supporting detail and examples. • Can develop a clear argument, expanding and supporting his/her points of view at some length with subsidiary points and relevant examples*. • Can evaluate the advantages and disadvantages of various options. • Can clearly signal the difference between fact and opinion.	• Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse. Though there may be some 'jumpiness' in a long contribution. • Can produce text that is generally well-organised and coherent, using a range of linking words and cohesive devices. • Can structure longer texts in clear, logical paragraphs.	• Can pass on detailed information reliably. • Can communicate the essential points even in more demanding situations, though his/her language lacks expressive power and idiomaticity.

PLURILINGUAL COMPETENCE		
Building on pluricultural repertoire • Can discuss the objectivity and balance of information and opinions expressed in the media about his/her own and other communities. • Can identify and reflect on similarities and differences in culturally-determined behaviour patterns (e.g. gestures and speech volume) and discuss their significance in order to negotiate mutual understanding. • Can, in an intercultural encounter, recognise that what one normally takes for granted in a particular situation is not necessarily shared by others, and can react and express him/herself appropriately. • Can generally interpret cultural cues appropriately in the culture concerned. • Can reflect on and explain particular ways of communicating in his/her own and other cultures, and the risks of misunderstanding they generate	*Plurilingual comprehension* • Can use his/her knowledge of contrasting genre conventions and textual pattern in languages in his/her plurilingual repertoire in order to support comprehension.	*Building on plurilingual repertoire* • Can alternate between languages in his/her plurilingual repertoire in order to communicate specialised information and issues on a subject in his field of interest to different interlocutors. • Can make use of different languages in his/her plurilingual repertoire during collaborative interaction, in order to clarify the nature of a task, the main steps, the decisions to be taken, the outcomes expected. • Can make use of different languages in his/her plurilingual repertoire to encourage other people to use the language in which they feel more comfortable

[1]*This category does not distinguish between B2.1 and B2.2.*

[2]*North, Goodier, and Piccardo 2018, 107: "In the two scales, Language A and Language B may be two different languages, two varieties of the same language, two registers of the same variety, or any combination of the above."*

Appendix B Transcript Conventions

The following transcript conventions are taken from Richards 2003, 173–74. Courier font is used as a monospaced typeface to ensure visual alignment of transcribed text. Speakers in the third column are indicated as IR (= interviewer) and IE (= interviewee). Examples are drawn from the interview transcripts, some with slight modifications.

Symbol	**Description**	**Example**
.	Falling intonation	It depends.
,	Continuing contour	Absolutely, it might be a good idea
?	Questioning intonation	Would that matter?
!	Exclamatory utterance	Yeah!
(2.0)	Pause of about 2 seconds	Uh, (2.0) ever?
(. . .)	Pause of about 1 second	It will maybe, uh, make a (. . .) difference.
(..)	Pause of about 0.5 second	I think that's an interesting (..) point.
(.)	Micropause	Say we marked an essay (.) at 78%,
´`   `´	Overlap	IR I want to switch over to the IB side, then- ´like,` IE ˛ Juu!˛
⟦	Speakers start at same time	⟦IR .hh and- ⟦IE Yeah, so it's-
=	Latched utterances	IR: Yeah, they do= IE: =which we don't.
__	Emphasis	The evaluation is <u>primarily</u> based on the first criterion.
:	Sound stretching	This is wa::y more academic.
(xxx)	Unable to transcribe	(xxxxx) was the best in her class, I think.
(send)	Unsure transcription	No. So many (tabs open).
(( ))	Other details	they didn't necessarily do it very well like ((laughs)) hhh you know
↑	Prominent rising intonation	I think it was well ↑ structured.
↓	Prominent falling intonation	What ↓ I like about the IB assessment criteria
-	Abrupt cut-off	to- to some extent, but the content
(x)	Hitch or stutter	(x) It's closer than the other ones.
CAPS	Louder than surrounding talk	NOT JUST ON THE GROUND, but also the examination board
hhh	Aspirations	Hard to say. hhh Hard to say.
.hhh	Inhalations	I mean, .hhh it needs to be there for transparency
° °	Quieter than surrounding talk	we look at, uh, °natural sciences°
> <	Quicker than surrounding talk	cause it's >easier to compare with paper one<
< >	Slower than surrounding talk	specifically about the <essay or course contents?>

Appendix C Interview Guide

Informants were given access to a Google Drive, accessible only by City of Espoo G Suite accounts. Files therein were prepended with a number, as referred to in question 8. Informants are referred to as T##, based on the number assigned to them upon anonymization of data.

1. How did going through the sample essay compare to one usually written for a first-year English course?
2. Did it feel natural to apply the Finnish matriculation exam criteria to the sample essay?
    a. IB: How familiar were you with these criteria?
    b. National: How familiar were you with the IB criteria?
3. How did the formulation of the criteria in the two curricula change the way you thought about the essay, if at all?
4. I find myself often relying on the formulation of the criteria rather than checking to see how students understand the marking criteria or my application of the marking criteria. Do you think that the IB or national matriculation exam criteria are helpful in the process of improving student writing?
5. (Comment on sample essay). How do you try to use these comments to help improve the student's writing?
    a. T01: Your comments and revisions focus largely on grammatical mistakes but also focus on structural issues, proper argumentation, and clarity. Compared to the examiner, your marks agreed exactly, except on understanding, in which criterion you marked the essay one point higher.
    b. T02: You used a very minimal method of correction by color coding the text and commenting on how individual sentences could be improved rather than suggesting revisions. Compared to the examiner, you were more generous when grading the content, equal on structure, but slightly harsher on the language.
    c. T03: You offer a lot of positive feedback alongside constructive criticism, focusing especially on structure and clarity of argumentation. Compared to the examiner, you were much more generous on the first two criteria, awarding 2 more points in both; your evaluation otherwise agreed with that of the examiner.
    d. T04: You've used the same color-coding system as the other national-side teachers and use comments both to suggest direct corrections to the text and less direct corrections and what should be fixed in the language or structure. Some comments even ask questions to try to provoke self-realization, like "How can you punctuate this better to help the reader?" Compared to the examiner, your overall assessment was in complete agreement, though you were slightly more critical of the language/structure and slightly more lenient on the content of the essay.
    e. T05: You comment extensively on the essay without directly correcting almost anything and seem to offer Socratic questions to try to prompt self-realization about stylistic infelicity and grammatical errors. Compared to the examiner, your evaluation was in agreement regarding language/structure, but you were more lenient on the content of the essay.
6. Do you think this kind of task successfully assesses a students' performance in a real-life situation? What would that real-life situation be?
7. Is your consideration of students' real-life use of English change impacted by the fact that high school students are likely to attend university?
8. Please scan through the exams for written language production (05) in the Drive folder. Do you think that one approach better assesses real-life language production and interaction than the other?
    a. Do you think that there are higher demands for formality in language production in one exam over the other? What other differences do you find interesting?
    b. How familiar were you with how written production is assessed in the IB/national curriculum?
    c. Are the exams an accurate reflection of written English-language tasks in your curriculum (nat: 2015 / IB: 2019)? If not, how else is real-life written language production developed? How do these tasks link to the assessment of written language production?

9. How active is your role as a teacher for written language production? Do you see yourself as a leader, advisor, model, facilitator, co-communicator, coach—a combination of these roles?
10. What balance do you think should be struck between grammatical accuracy, personal style, praise, ability to communicate, and appropriateness to social context when assessing and developing student written production?
11. With the few remaining questions, I want to turn to the essay scores gathered for first-year students in school year 2019–2020 and combining English-language instruction for school year 2021–2022. Do you find when assessing written language production that one gender outperforms the other or that you tend to favor writing by one gender?
12. How do you think the different mother-tongue backgrounds of [the school]'s students impact on their capacity for and ability to develop in written language production?
13. Does your assessment of student written language production change when assessing first-year students as opposed to older students?
14. How much consideration do you think should be given to real-life written language production?
15. What about, more specifically, in terms of *academic* written language production?
    a. Would you would consider this real-life written language production?
16. What challenges do you think need to be addressed for combining the instruction and assessment of written language production for first-year students at [the school]?

Appendix D Coding Themes

The following coding themes were developed primarily from CEFR and the interview transcripts, with other primary source material (i.e., curricula, online learning platforms, exam criteria) used to extract further themes.

Coding Theme	Sub-type	Code
Assessment	1	Peer
		Self
		Teacher
	2	Formative
		Summative
	3	Oral
		Written
Validity	N/A	Test-Taker Characteristics / Curriculum
		Cognitive
		Context / Task Familiarity
		Scoring
		Criterion-related
Communicative Language Activities and Strategies	Production, activity	Creative Writing
		Essays and Reports
	Production, strategy	Compensating
		Monitoring and Repairing
		Planning
	Mediation, activity	Communication
		Concepts
		Texts
		Explaining
		Simplifying
	Reception, activity	Audio-visual
		Listening Comprehension
		Reading Comprehension
	Reception, strategy	Cues
	Action orientation	Social context
Communicative Competence	Linguistic	Sentence Level
	Sociolinguistic	Register
		Style
		Politeness
	Pragmatic	Thematic development
		Coherence and Cohesion
	FNC term	Communicativity
	Sample Essay	Sentence-level Errors
		Above-sentence-level Errors
Plurilingual and Pluricultural Competence	N/A	Building on Pluricultural Repertoire
		Plurilingual Comprehension
		Building on Plurilingual Repertoire
Approaches to Teaching	N/A	Backwash
		Differentiation
		Classroom Management

		Explicit or Implicit Instruction
		Source Text
		Task Types
		Teacher Role
Approaches to Learning	N/A	Informal
		Language Profile
		Metacognition
		Misconceptions
		Special Needs
		Transfer
		Classroom Dynamics
Affect	N/A	Student
		Teacher
Challenges and Opportunities	N/A	Challenges
		Opportunities

Appendix E Essay Sample

The essay and evaluation form used in the experiment with the teachers is reproduced below.

EXAMPLE PAPER

Diploma Programme
Programme du diplôme
Programa del Diploma

English A: language and literature – Standard level – Paper 1
Anglais A : langue et littérature – Niveau moyen – Épreuve 1
Inglés A: lengua y literatura – Nivel medio – Prueba 1

1 hour 15 minutes / 1 heure 15 minutes / 1 hora 15 minutos

Instructions to candidates

- Do not open this examination paper until instructed to do so.
- Answer one of the two questions.
- The maximum mark for this examination paper is **[20 marks]**.

Instructions destinées aux candidats

- N'ouvrez pas cette épreuve avant d'y être autorisé(e).
- Répondez à l'une des deux questions.
- Le nombre maximum de points pour cette épreuve d'examen est de **[20 points]**.

Instrucciones para los alumnos

- No abra esta prueba hasta que se lo autoricen.
- Responda una de las dos preguntas.
- La puntuación máxima para esta prueba de examen es **[20 puntos]**.

4 pages/páginas

2216 – 2015

International Baccalaureate®
Baccalauréat International
Bachillerato Internacional

Write an analysis on one text only. It is not compulsory for you to address the guiding question in your answer.

TEXT A

[20 marks]

NATURE | EDITORIAL

A question of time

Timekeeping is boosted by the advent of an optical clock based on strontium atoms.

22 January 2014

PDF Rights & Permissions

When the history of the twenty-first century comes to be written, one of the most puzzling questions asked will be why, well into the information age, millions of people still paid to dial a number on their phone to find out the time. Almost 80 years after its formation, the UK speaking clock, the world's original telephone time service, remains an essential part of British life. This is despite the near ubiquity of time displays — not least on the mobile phones that people discard to call 123 from a fixed line.

For some people, at some times, accuracy matters. Peaks in the use of the speaking clock come, for instance, on New Year's Eve, or when the clocks are put forward and back by an hour to mark, respectively, the start and end of British Summer Time.

Related stories

- An optical lattice clock with accuracy and stability at the 10−18 level
- Precise atomic clock may redefine time
- Atomic clocks use quantum timekeeping

There is another way, at least in Britain. BBC Radio regularly broadcasts the same time signal used to set the speaking clock — affectionately known as the pips. Indeed, it has become as much a feature of some shows as the content planned around it. Time is more than a British institution; it is woven into the cultural fabric of everyday life.

The pips are drawn from an atomic clock held at the National Physical Laboratory (NPL) in Teddington, near London. One of the most accurate in the world, the NPL clock is tuned to the regular bursts of light emitted by caesium atoms when they are excited by microwaves. The clock would lose roughly one second every 138 million years — a sufficient degree of accuracy for a bleary-eyed hour-late commuter who forgot to set their clock the night before, but not accurate enough for some.

In a paper published on *Nature*'s website this week, time lords in the United States describe the latest

advance in chronometry, and one that is as superior to the atomic pips as those pips were to the mechanical devices they replaced (B. J. Bloom *et al. Nature*http://dx.doi.org/10.1038/nature12941; 2014). The researchers have built a timepiece based not on caesium but on strontium. More importantly, it uses much higher, optical frequencies. This gives such devices, called optical clocks, greater accuracy than those that rely on microwaves. The new optical clock, for example, would not lose one second even if it were to run for 5 billion years.

It is also extremely stable — another key measure of timekeeping. (Accuracy defines how closely a clock's output matches the desired time signal, whereas stability is a measure of how steady that output is. A clock that loses precisely one second each day is inaccurate but stable, for example.)

The unveiling of the super-accurate strontium optical clock comes just a few months after a related group revealed a device based on ytterbium. Other laboratories across the world have their own designs. Inevitably, the increased precision and reliability of optical clocks are fuelling debate about whether they could be used to set the ultimate time, and redefine the second. (There are no official plans to do so, but plans are afoot to redefine other SI units.) These are heady times for metrology: a World View on page 455 describes attempts to measure another fundamental constant: Big *G*.

Nature has a particular stake in the race to develop new atomic clocks. Back in January 2003, we published a News Feature that surveyed the scene and tried to predict what would happen (D. Adam *Nature* **421,** 207–208; 2003). Within a decade, the piece suggested, optical clocks could rise to prominence and raise fresh debate about the definition of the second. A ten-year event horizon is a staple of scientific journalism, and most promised breakthroughs fail to materialize on deadline. The latest development in atomic timekeeping, by contrast, has arrived bang on time. Well, almost.

From the academic journal *Nature* Vol 505
(22 January 2014)

In what ways does the use of language in this article help to interest and entertain the reader?

Student Essay Sample

Note: The following text is meant to be a diplomatic transcription of a handwritten sample essay provided by the IB to help set the criteria. The IB examiner's marks will be discussed alongside yours during the interview. Please treat the essay as one written as practice for the matriculation/IB exam (under exam time constraints and being handwritten) by a first-year high school student in the third English course. ***Please mark and comment on the essay as you would normally (e.g., using the "suggesting" mode [View > Mode > Suggesting] and inserting comments), the aim being to foster growth in the student's written language performance****. Please evaluate the essay using both the IB and Finnish matriculation exam criteria (see below). Spelling and grammar suggestions are meant to be turned off, so please disable them in case they still appear (Tools > Spelling and grammar > uncheck "Show spelling/grammar suggestions").*

Text 1

The use of language can twist the mind of the reader, only if used right. What makes a text stand out in the eyes of the audience is its ability to send a message through its use of stylistic features and devices. Language can contain such tools to trigger, attract and influence readers. This improves the reputation of the creator and assumes that those readers will come back for another. The use of language to entertain the reader is done by the use of several techniques to address sertain issues. These techniques include the struction and placement of content, stylistic features used and the type of audience the text is dedicated to. The Article "A question of time" by Nature uses several techniques to make sure theri audience remain entertained as they reach the end of the article.

"Nature" is a website for an international audience to receive scientific information. Although it addresses the British culture frequently, it is dedicated to cultures all over (line 14). It gives examples of large and popular events such as "New Year's Eve" and known businesses such as the BBC Radio. This targets fimiliarity with such brands and allows the reader to feel understood like this article was made for them. This attracts a wide range of audiences specially since such ideas are known all over the world.

The details of the content are listed in a way to build the argument. Starting of with a calm maroon color on the display for the websites title. The color is earthy showing how grounded and comferting the website is. If bright colors where used the reader will be overwhelmed by the appearence specially since there are alot of words displayed on the page. The title is bolded and wide (almost as wide as the website title) to capture the attention of the reader first. Followed with a short description on what the article will be discusing to inform the reader what to expat. The article starts with a time lapse from a point in history up till now. The use of short paragraphs allow the reader to enjoy and not to rush reading them, specially with the short sentences used. A "related stories" option here is placed next to the paragraphs that contain huomur and known examples to show the reader that "there is more if you want". This exites the reader and drives them into reading the rest. In addition to the articles by "Nature" that where referenced for readers to go to as they read the text. All of these are used to the main topic this

text is about. Which is described clearly and effectively. The audience are not lost and their reactions build up as the argument and the texts build up. Overmore, the listing of 3's. This is used to grasp on the reader by giving evidence in the list of 3s and by showing the wide varity of choices (line 6).

The tone used is gentle yet informative. Easy language the reader doesn't have to think too much about and can continue reading. In addition to the use of words taht link ideas together including sentence startups such as "More importantly" (line 23) that not only indicates that the previous information is important but there is ever more important information. Specially when the voice of the article is an educated, well-informed person. The exitement comes fro mthe words that the audience are spose to feel. The use of the word indeed (16) to not only asure the reader of the information but also to keep them reading on what is it that is 'indeed' since its placed at the start of the sentence. In addition, the use of generalisation to make the reader feel the importance of the article such as when mentioning "one of the most puzzling questions asked" (line 1). Including the audience on a global scale is also a form of entertainment. Not only because it includes everyone from everywhere but also because no-one wants tobe left out since humans are social creatures that work the best in packs and being discluded from the world would be a nightmare for some.

This texts contains stylistic features that inhances its ability to hold on to the reader. Some include its its open endings for each paragraph. This connects the reader to the next one as soon as they thought that the ideas was over, such as "but not accurate enough for some" (line 19) after mentioning scientific information. Changing sentances was also used for the last sentence which contained only two words. This feeling of uncertainty leaves the reader with more questions that might lead him to one of the "Related Stotires" box of options or one of the articles refrenced in the text.

The uses of all of these devices to inform the reader about scientific knowledge but also to entertain them. Although the use of some scientific vocab such as elements (caesium (line17) might seem boring and difficult for some, the author explained them so that they can be understood for people all over the world.

Some people tend to stay away from certain subjects because of their bias perspective of how difficult it might seem. The use of an open title for the article lights the curiosity of the reader and once they start reading and understanding, language grasps and hold on to them till the end, specially with an earthy and cool atmosphere the content is placed on. Language can twist our minds. And we fall for it.

Evaluation

Note: Please include a brief rationale for each criterion, as well as an overall comment. One to two sentences should suffice for comments on each criterion. Please write the overall comment however you would normally, keeping in mind that this essay would have been submitted as a handwritten exam at the end of a student's first year.

Grade according to IB Criteria

Criterion	Max. Points	Points Awarded	(Brief) Rationale
Understanding and Interpretation	5		
Analyis and Evaluation	5		
Focus and Organization	5		
Langauge	5		
Overall	**20**		

Grade according to Finnish Matriculation Exam criteria

Criterion	Max. Points	Points Awarded	(Brief) Rationale
Communicativity	99		
Meaning and Structure	99		
Breadth and Accuracy of Langauge	99		
Overall (based primarily on "Communicativity")	**99**		

Appendix F Research Permission Form

The research permission form for the master's thesis study is reproduced below, granted by the City of Espoo (5 May 2020), with my personal information and the name of the school redacted.

ESPOO
ESBO

Espoon sivistystoimi Tutkimuslupahakemus

1. TUTKIMUKSEN NIMI	Action-Oriented Language Production in the Assessment of EFL Writing in the Finnish High School Curriculum and the International Baccalaureate Diploma Program: A Case Study of a High School in Espoo, Finland	
2. KOHDEYKSIKKÖ	Tutkimuksen suunniteltu kohdeyksikkö (-yksiköt) Espoon kaupungissa [redacted]	
3. TUTKIMUKSEN KUVAUS	Lyhyt kuvaus tutkimuksen sisällöstä ja menetelmistä (max. 160 merkkiä). (Liitä tutkimussuunnitelma liitteeksi.) Tutkimus tarkastelee, miten englannin kielen opetus IB- ja kansallisessa opetussuunnitelmassa heijastaa toimintaorientoitunutta kielen opetusta. Onko tutkimussuunnitelma salassa pidettävä? ☐ Kyllä, perustelut: ☒ Ei, tutkimussuunnitelma on julkinen.	
	Aineiston otanta ja keruuaika Opettajien haastattelut toteutetaan toukokuussa 2020, ja esseearvosanat kerätään kesäkuussa 2020.	
	Tutkimuksen tarkoitus ☒ Pro gradu ☐ Lisensiaattityö ☐ Väitöskirja ☐ Muu opinnäytetyö, mikä ☐ Muu, mikä? Tutkimuksen arvioitu valmistumisaika: Marraskuu 2020	
4. TUTKIMUKSEN TOTEUTUS	Ovatko tutkimuksen kohteena sivistystoimen asiakkaat, esim. oppilaat?	☐ Kyllä ☒ Ei
	Onko tutkimuksen kohteena sivistystoimen henkilöstö?	☒ Kyllä ☐ Ei
	Onko tutkimuksen kohteena henkilö (henkilöitä), jonka osallistumisesta päättää huoltaja tai edunvalvoja	☐ Kyllä ☒ Ei Jos kyllä, selvitä Lisätietoja -kohtaan, miten huoltajan suostumus hankitaan
	Tutkittavien henkilöiden lukumäärä	5
	Käsitelläänkö tutkimuksessa henkilötietoja	☒ Kyllä ☐ Ei
	Muodostuuko tutkimusta tehtäessä henkilötietopohjainen tutkimusrekisteri	☒ Kyllä ☐ Ei Jos kyllä, täytä myös Tutkimusrekisteritiedot -lomake
	Käytetäänkö tutkimuksessa jo olemassa olevien rekistereiden tietoja	☐ Kyllä ☒ Ei Jos kyllä, selvitys Lisätietoja -kohtaan
	Onko tutkimus osa jotain laajempaa tutkimusta / projektia	☐ Kyllä ☒ Ei Jos kyllä, selvitys Lisätietoja -kohtaan.
	Aineiston keruumenetelmä	☐ Kysely ☐ Havainnointi ☒ Haastattelut ☐ Asiakirja-analyysi ☒ Muu, mikä: arviointi-analyysi

	Miten tutkimusaineisto säilytetään tietoturvallisesti tutkimuksen teon ajan (esim. lukollinen kaappi, salasanat, kulunvalvonta, käyttöloki, pseudonymisointi)? salasanasuojattu tiedosto
	Tutkimusaineiston hävittäminen tai arkistointi tutkimuksen päätyttyä ☐ Tutkimusaineisto ja tunnistetiedot hävitetään. Miten aineisto tuhotaan tietoturvallisesti ja milloin? ☒ Tutkimusaineisto arkistoidaan ilman tunnistetietoja. Miten tunnistetiedot tuhotaan tietoturvallisesti ja milloin? Tunnistetiedot tuhotaan koneelta anonymisoinnin jälkeen. Anonymisoitu data julkaistaan osana tutkimusta e-thesiksessä. ☐ Tutkimusaineisto arkistoidaan tunnistetiedoin tutkimuksen päätyttyä arkistolain mukaisesti. Miten tutkimusaineisto arkistoidaan tietoturvallisesti ja minne?
	Lisätietoja
5. TUTKIJATAHON TIEDOT	Tutkimuksen tekijä/t (alleviivaa yhteyshenkilö) Kenneth Wenchen Lai
	Yhteyshenkilön osoite [redacted] | Puhelin [redacted]
	Sähköpostiosoite kenneth.lai@espoo.fi
	Organisaatio / yksikkö, johon tutkimus tehdään English Studies- maisteri ohjelma, Helsingin yliopisto
	Tutkimuksen ohjaaja / vastuullinen johtaja yhteystietoineen Tuomo Hiippala, tuomo.hiippala@helsinki.fi

6. TUTKIMUKSEN HYÖDYT	Arvioi, miten tutkimus hyödyttää kaupungin palvelujen kehittämistä: Tutkimus auttaa yhdistämään ensimmäisen vuoden opiskelijoiden IB-linjan ja kansallisen puolen opiskelijoiden opetusta LOPS2021 varten.
7. TUTKIMUKSEN TEKIJÖIDEN SITOUMUS JA ALLEKIRJOITUKSET	Vakuutan, että tässä tutkimuslupahakemuksessa ja sen liitteissä annetut tiedot ovat oikeat. Sitoudun siihen, että en käytä saamiani tietoja tutkimuksen kohteen tai hänen läheistensä tai Espoon kaupungin vahingoksi tai sellaisten etujen loukkaamiseksi, joiden suojaksi on säädetty salassapitovelvollisuus. En luovuta saamiani henkilötietoja sivullisille, vaan pidän ne salassa. Tutkimustulokset esitän niin, ettei niistä voida tunnistaa yksittäistä henkilöä tai perhettä. En käytä saamiani tietoja muuhun tarkoitukseen kuin mihin tutkimuslupa on myönnetty. Noudatan EU:n yleistä tietosuoja-asetusta, tietosuojalakia ja muualla lainsäädännössä mainittuja säännöksiä henkilötietojen käsittelystä ja salassapidosta. Sitoudun tutkijan eettisiin periaatteisiin, tutkimuksen toteutusehtoihin ja sivistystoimen antamiin ohjeisiin. Ilmoitan viipymättä tutkimuslupahakemuksessa pyydettyjen henkilötietojen tietoturvaloukkauksesta Espoon kaupungin tietosuojavastaavalle

	tietosuoja@espoo.fi, jos tutkimusryhmän muu jäsen ei ole vielä ilmoitusta tehnyt. Suostun siihen, että Espoon kaupungin internet-sivuilla julkaistaan tutkimuksen nimi, tutkimuksen tekijän organisaatio ja tutkimuksen arvioitu valmistumisaika.
	Paikka ja aika Helsinki, 27.4.20
	Allekirjoitukset ja nimenselvennykset Kenneth Wenchen Lai
8. PÄÄTÖS	☒ Tutkimuslupa myönnetään ☐ Tutkimuslupa myönnetään ehdollisena: Myönnetyn tutkimusluvan numero /20 ☐ Tutkimuslupaa ei myönnetä seuraavin perustein:
	Pyydetään lähettämään tutkimuksen valmistuttua sähköpostitse samaan osoitteeseen kuin tämä tutkimuslupahakemus ☒ Tiivistelmä ☐ Koko tutkimusraportti
	Espoossa 5 / 5 20 20
	Päättäjän allekirjoitus
	Nimenselvennys Ida Stolt-Haglund
	Virka-asema kehittämispäällikkö
Tutkimusluvan myöntäminen ei velvoita tutkimuksen kohteita osallistumaan tutkimukseen. Tutkijan on neuvoteltava aina erikseen tutkimuskohteena olevien organisaatioiden kanssa tutkimukseen osallistumisesta ja kohteen nimen mainitsemisesta tutkimusraportissa. Tutkimuksen teko ei saa häiritä tutkimuskohteen toimintaa.	

9
LIITTEET Merkitse alle rastilla

- ☒ Tutkimussuunnitelma
- ☐ Tutkimusrekisteritiedot
- ☐ Haastattelurunko/kyselylomake
- ☐ Suostumuslomake
- ☐ Tiedote/tiedotteet tutkimuksesta
- ☐ EU:n yleisen tietosuoja-asetuksen mukainen vaikutustenarviointi
- ☐ Muu, mikä?

Appendix G Online Data Repository

The non-anonymized data used for this thesis (i.e., excluding the video recording of the interviews and the essay score data set built using student names), as well as the statistical and graphic-producing procedures executed in SPSS, can be found at and downloaded from the following link: https://doi.org/10.5281/zenodo.4460549.